Evaluation and analysis of stepped wedge designs

Zhan, Zhuozhao

EVALUATION AND ANALYSIS OF STEPPED WEDGE DESIGNS

APPLICATION TO COLORECTAL CANCER FOLLOW-UP

Zhuozhao Zhan

ISBN	978-94-034-0422-6 (printed version) 978-94-034-0421-9 (digital version)
Cover design:	Zhuozhao Zhan
Lay-out:	Zhuozhao Zhan
Printed by:	Ridderprint BV, Ridderkerk
LaTeX template:	Based on the LaTeX PhD thesis template of Technical University Delft

Financial support for the publishing of this thesis was kindly provided by University of Groningen, University Medical Center Groningen, and Research Institude SHARE.

An electronic version of this dissertation is available at
http://www.rug.nl.

rijksuniversiteit
groningen

EVALUATION AND ANALYSIS OF STEPPED WEDGE DESIGNS

APPLICATION TO COLORECTAL CANCER FOLLOW-UP

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken,
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

maandag 12 februari 2018 om 14.30 uur

door

Zhuozhao Zhan

geboren op 10 november 1988
te Zhejiang, China

Promoteres
Prof. dr. G. H. de Bock
Prof. dr. E. R. van den Heuvel

Beoordelingscommissie
prof. dr. H. M. Boezen
prof. dr. G. J. P. van Breukelen
prof. dr. G. Beets

Paranimfen
Anne Looijmans
Xuan Anh Phí
Bronislaw Abramiuc

CONTENTS

1

INTRODUCTION

1

1.1. CHAPTER OUTLINE

The randomized controlled trial or randomized clinical trial is considered the gold standard for establishing efficacies and effectiveness of a new intervention in the medical field. In a randomized controlled trial, participants are randomly allocated to two or more treatments and data is collected from those treatment arms for comparison. A randomized controlled trial ensures that participants in different treatment arms only differ in terms of the treatment they receive and therefore difference in the outcomes can be attributed to the difference in treatments.

Randomization can be performed either at an individual level or at a cluster level. A randomized controlled trial with randomization at a cluster level is often named a clustered randomized trial. The advantage of randomization at a cluster level over randomization at an individual level is that it protects the trial from possible contamination, and that it is easier to perform when individual randomization is not feasible. However, it is more difficult to maintain balance in possible confounders for different treatment arms since participants from the same clusters will be assigned to the same treatment.

In a randomized controlled trial different clinical trial designs may be applied, including the well-known parallel group design. Other types of clinical trial design such as crossover design, factorial design are also frequently applied. Randomized controlled trials may also apply a sequential introduction of a new treatment to determine its efficacy or effectiveness with respect to a control treatment. In contrast with the classical parallel group design where different treatments are assigned to distinct groups of patients, sequential introduction of the treatment usually applies the two different treatments to the same (group of) patients or clusters in a chronological order. Such design provides opportunities for within-subject or within-cluster comparisons in addition to the between-subject

or between-cluster comparison, creating a design where patients or clusters could be considered as their own control. This type of clinical trial design is the so-called stepped wedge design.[16] This will be the topic of this thesis entitled "Evaluation and analysis of stepped wedge designs: Application to colorectal cancer follow-up" in which the epidemiological practice and application of the stepped wedge design will be discussed.

The arguments for the application of a stepped wedge design, factors to consider when designing a trial using a stepped wedge design, and the statistical analysis of data obtained from a stepped wedge design will be demonstrated on the basis of the CEA-Watch study.[32] The CEA-Watch study is a clinical trial investigating an intensification of the follow-up protocol for colorectal cancer as compared to care as usual follow-up in patients after surgical resection of their primary tumor. In the introduction to the present thesis, several aspects of the stepped wedge design and the CEA-Watch study will be introduced and discussed briefly. The chapter starts with a general introduction to the stepped wedge design, its primary merits and limitations, followed by a short historical overview of the application of the stepped wedge design in the field of cancer epidemiology and some descriptions of the CEA-Watch study. Afterwards, the application of the stepped wedge design to the motivating example, i.e., the CEA-Watch study will be outlined. This chapter will end with the aims and outline of this thesis.

1.2. STEPPED WEDGE DESIGN

A stepped wedge design is a randomized controlled trial design that utilizes sequential roll-out of the intervention. [6] At the beginning of the trial, a patient or group of patients start within the control or placebo arm. Switching to the new intervention of interest then take place at predetermined moments. During each switch moment, a part of the control

1

arm crossover to the intervention arm. There are no switches from the intervention arm to the control arm. In the last time period of the trial, namely the period between the last switch moment and the end of the trial, all patients or group of patients are in the intervention arm, and exposure is exclusively to the intervention. An example of a stepped wedge design with three switch moments and two hospitals per switch moment is depicted in Figure 1.1 below.

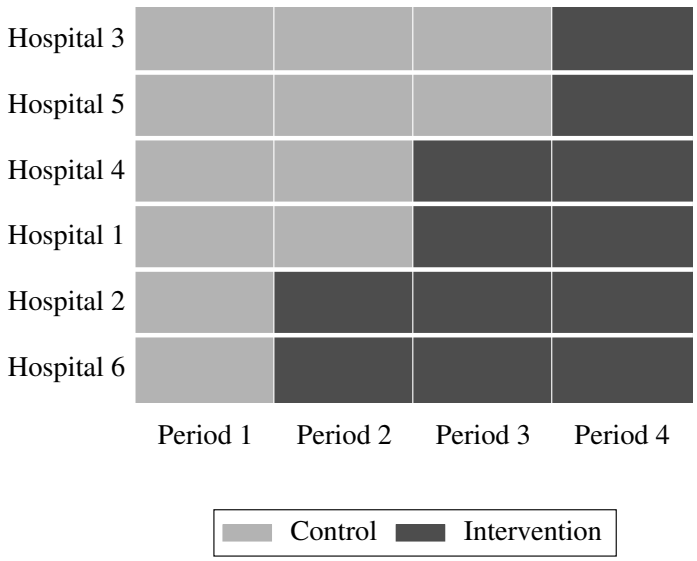

Figure 1.1 | A schematic of a stepped wedge design with three clusters (two hospitals per cluster) and four periods (light periods indicate control and dark periods indicate intervention)

Randomization in the stepped wedge design is used to allocate (groups of) patients to different switch moments instead of different treatment arms. Current literature on stepped wedge design is predominated by the clustered stepped wedge trial and randomization is conducted at the cluster level (Figure 1.1). The taxonomy of the stepped wedge design is based on the type of cohorts involved in the trial: a stepped wedge

design is considered to be cross-sectional if new patients are recruited and outcomes are measured step by step; on the other hand, if a static cohort is being followed throughout the course of the trial, this type of stepped wedge design is called longitudinal/cohort stepped wedge design; a combination of the cross-sectional and longitudinal stepped wedge design is named open cohort stepped wedge design.[15]

Though there are several reasons for adopting a stepped wedge design, there are three major motivations.[18, 27, 28] First of all, a stepped wedge design provides logistic conveniences and flexibilities when implementation of the new intervention cannot be realized simultaneously across all trial clusters. Such problem frequently rises from large-scale pragmatic trials [29] such as multi-center trials. To name a few causes of the difficulty, setups of the new intervention or learning the new techniques requires certain amount of time, or administration approval procedure needs different amount of time to be acquired among different locations. Under these circumstances, a stepped wedge design becomes particularly attractive since the new intervention does not need to be deployed concurrently. The second most common motivation is ethical considerations when withholding a new treatment for part of the cohort is considered unacceptable. This argument is even stronger when the efficacy of the treatment has already been demonstrated and proven. On the other hand, the ethical benefit is only true for a longitudinal or open cohort stepped wedge design, where all patients will eventually be exposed to the new intervention treatment. However, the cross-sectional stepped wedge design does not have this benefit since patients in a cross-sectional stepped wedge design adhere the same treatment as the treatment of their enrolled time period and cluster. Nonetheless, a consequence of the second motivation is the attraction of more participants and accessibility to a larger sample size. Last but not least, the stepped wedge design also provides statistical efficiency under certain conditions. For instances, it has

been shown that a stepped wedge design is more efficient in terms of estimating the treatment effect compared to the classical parallel group design when the intraclass correlation is large. An intuitive explanation is that, in a stepped wedge design, patients can be considered as their own control and therefore a stepped wedge design reduces the variations.

1.3. SAMPLE SIZE CALCULATION

Sample size calculation for stepped wedge design shares much similarity with traditional clustered randomized controlled trial.[8, 9] In this approach, the required sample size for a parallel group design will be estimated and the estimated sample size will be multiplied by the design effect of the stepped wedge design to obtain the required sample size. For the stepped wedge design, the design effect is the ratio of the treatment effect estimator variance of the stepped wedge design and the parallel group design. This approach is model dependent, as for different models and assumptions, the design effect will differ. In current literature, the design effect is available for cross-sectional and longitudinal stepped wedge design with certain variance components model for normally-distributed outcomes.[12, 14, 20, 34] An alternative approach is to directly calculate the variance of the test statistic and obtain the required sample size assuming the test statistics is asymptotically normal.[19] Nevertheless, both sample size calculation approaches rely heavily on obtaining the variance of the estimator or test statistic. For complex models and non-normal outcomes, this may not be feasible and options are limited to simulation-based calculation [2] except for binomial outcomes for which approximation by normal distribution may be adopted.

1.4. STEPPED WEDGE DESIGN IN CANCER EPIDEMIOLOGY

The root of using stepped wedge design in cancer epidemiology traces back to early 1990's. In the 1983 report of WHO meeting on the prevention of liver cancer [35], several points on designing large-scale studies to evaluate the effectiveness of immunization in preventing hepatocellular carcinoma were mentioned. It was suggested that ethical problems would be present for traditional randomized controlled trial design since vaccination of children with hepatitis B vaccine at birth would confer long-term protection against the development of hepatocellular carcinoma. Another problem was related to the costs and limited supplies of the vaccine. It was reported that for a potential trial at multiple sites, it would be unlikely to have sufficient vaccine available for all sites, especially for high-risk regions such as Africa and Asia. But considering the long follow-up time of such trial, it might be the case that the vaccine would become cheaper and more widely available during the course of the trial. Motivated by these problems, the so-called "Gambia Hepatitis Intervention Study" was conducted in The Gambia in 1987 [11] which is considered the earliest stepped wedge design trial recorded in literature. In this study, the hepatitis B virus vaccination was introduced to the "Extended Program of Immunization" in Gambia at approximately 10- to 12-week intervals by vaccination teams. All new born children recorded at the vaccination points served by the team was included. Vaccination effect was evaluated through a long-term follow-up of these children during the trial period and incidence of hepatocellular carcinoma and chronic liver diseases was compared among vaccinated children and those who were not in each 3 months period.

More recent accounts for stepped wedge design used in cancer epidemiology related trials, started to appear after the seminal paper from Hussey and Hughes in 2007 [20] and the resurface of stepped wedge design

in clinical trials in general.[3] In a review paper on breast cancer screening [10], it was suggested that multiple time series data might be particularly useful in evaluating screening introduced across health systems or countries at different times. The stepped wedge design was proposed as one of the randomized trial design options. In the same days, trials with a focus on psychological outcome in cancer patients often used pre/post comparisons within a randomized controlled trial framework.[17, 26] This type of evaluation method (or its variation) can also rise from a stepped wedge design when measurements of psychological variables are performed at multiple time periods during the trial. Essentially, pre-/post comparison can be viewed as one form of the stepped wedge design by using one single switch moment. This has also been suggested by a systematic review on improving quality of care for lung cancer.[36] Other topics of cancer-related trials which used the stepped wedge design vary from medical education for general practitioners [31], healthcare for cancer patients [1, 4, 5], to community-based care support to primary colorectal cancer diagnosis using immunochemical faecal occult blood test.[22] The two most frequent mentioned rationales for using a stepped wedge design are ethical considerations and logistic limitations which is consistent with the general motivation as outlined above.

1.5. THE CEA-WATCH TRIAL

In this thesis, the primary motivating example is the CEA-Watch study.[32] It is a multi-center clustered stepped wedge trial conducted in 11 non-academic teaching hospitals in the Netherlands between the period of 2010 and 2012. The objective of the study was to investigate whether intensification of the follow-up protocol would be associated with a higher percentage of early stage recurrences and an increased survival time in patients with recurrences as compared to care as usual follow-up protocol.

Thereby, developing an evidence-based guideline for routine follow-up using standard screening tools, namely CEA, short for carcinoembryonic antigen, and imaging techniques such as computer tomography. Eligible participants were colorectal cancer patients with American Joint Committee on Cancer (AJCC) stage I-III after R0 resection whom had been operated from 2007 until July 2012. Patients who were not medically fit for metastasectomy, diagnosed with other malignancies or had metachronous metastases at the start of the trial were excluded. The intervention follow-up protocol adhered to CEA measurements every two months and yearly imaging in patients' first three years of follow-up. Outpatient clinic visits with imaging of chest and abdomen were scheduled on a yearly basis in the same period. In case of more than 20% increase in CEA value with absolute value higher than 2.5 ng/ml, another blood sample was drawn. If a consecutive rise was observed, a CT scan of chest and abdomen was advised. The care as usual follow-up protocol following the 2008 national guideline of the Netherlands was considered as the control arm and is consisted of outpatient clinical visits every 3-6 months in the first three years of patients' follow-up and every single year in the last two years. A comparison between the two follow-up protocols is shown in Figure 1.2. The primary outcome of the trial was the percentage of curative treated recurrences among all patients which are considered at risk for developing recurrences. Secondary outcomes included time-to-detection of the recurrences, quality of life and mental wellness of the patients, and long term survival for patients with recurrences.

The eleven participating hospitals were grouped into five clusters, three smaller hospitals were grouped together to ensure clusters were balanced in terms of the number of patients. Every three months, a cluster switched from the care as usual protocol to the intervention protocol until all clusters were switched. The order of the switch is randomized using simple randomization. Substantial overlapping between the recruiting period

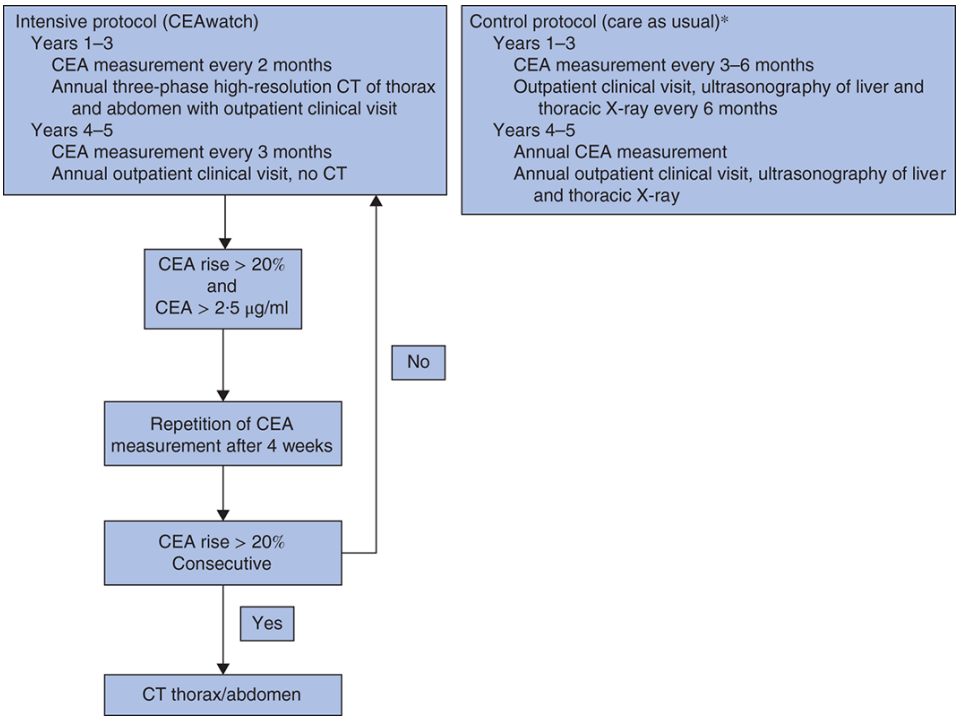

Figure 1.2 | Schematic illustration of the intensive carcinoembryonic antigen (CEA) measurement (CEAwatch) and control (care as usual) follow-up protocols. *Local differences and adjustment by individual hospitals allowed

and trial period indicates that new patients were included during the trial which makes the stepped wedge design an open cohort one.

1.6. STEPPED WEDGE DESIGN FOR CEA-WATCH

The rational for adopting a stepped wedge design in the CEA-Watch trial will be briefly discussed here. As the use of CEA for colorectal cancer follow-up have been established [7, 21, 23, 30, 33], the next step was to evaluate its effectiveness and cost-effectiveness on the community level. Randomized and controlled experimental approaches might have the highest internal validity, but are not always feasible and are difficult and

costly to implement in larger population to provide support for evidence-based decision making. Moreover, these are often not best suited for testing complex interventions with long term lifestyle changes.[25] As the CEA-Watch study promoted a systematic evaluation of a complex follow-up protocol on its pragmatic nature, stepped wedge design was appealing for the investigators when planning the trial. As a prerequisite, the CEA-Watch trial required a computer assisted system [13] to be installed and functioning at each participating hospital to ensure the validity and adherence of the follow-up under study. It was deemed unrealistic to start multiple hospitals with the new intervention at the same time. In addition, approvals from each local medical ethic committees were anticipated to be cumbersome. Thus the stepped wedge design became one of the better choices for its staggered starting points.

However, adopting a relatively new trial design such as the stepped wedge design also imposed challenges and problems. From a designing prospective, some of the limitations have already been foreseen during the planning phase. For instance, it could be expected that there would be an increased risk of attrition due to the prolonged waiting time for the new intervention as well as an increased risk of contamination. These potential problems put doubtful clouds above the trial validity and biasness [24], and therefore requires prudent examination. On the other hand, from a statistical or data analysis perspective, resources and information with regards to analyzing different endpoint outcomes under the contexts of the stepped wedge design are limited in literature. Thus far, only the method for analyzing continuous normally-distributed outcome has been addressed.[12, 20] It is unknown for other types of outcomes, such as relative risk or survival time, whether the traditional statistical methods would be still appropriate. If not, then what kind of adjustment is needed? Especially in the CEA-Watch trial, the primary outcome is a relative risk type of outcome and it also has survival time and questionnaire as its

secondary outcomes. Therefore, there is a need to investigate on these questions and demonstrate the proper methods.

1.7. THESIS AIMS AND OUTLINE

1.7.1. AIM

The aim of the thesis is to investigate the practice of the stepped wedge design for epidemiological studies, specifically large-scale pragmatic clinical trials. Furthermore, to discuss and illustrate the appropriate data analysis methods for various types of outcomes commonly seen from such trials.

1.7.2. OUTLINE

In the first part of the thesis, the focus is on the theoretical discussion of the stepped wedge design. In Chapter 2, its common strengths and weaknesses are surveyed and summarized. The merits of the application of a stepped wedge design specific to the CEA-Watch study is closely examined and discussed. The results demonstrate that not all perceived traits of the stepped wedge design apply to the CEA-Watch study. The implications from this chapter can be generalized to more situations similar to the CEA-Watch study. In Chapter 3, data analysis methods for the stepped wedge design are discussed. It demonstrates the usage of different methods for different outcome types and highlights the assumptions made by these methods, when to use or not use such method, and what are the caveats to consider when analyzing the data that arise from a stepped wedge design trial.

The second part of the thesis consists of three examples of data analyses from the CEA-Watch study. Chapter 4 considers the main outcome of interests, the proportion of recurrences with curative treatment as well as the time-to-detection of the recurrences. This proportion can be con-

sidered as a binary outcome and the time-to-detection is a survival time type of variable. The difficulty lies in the fact that patients switch treatment in the stepped wedge design so treatment need to be considered as time-dependent. In Chapter 5, the long term survival time is evaluated for patients that have developed recurrence during the trial. It is necessary to distinguish this with the survival time analysis showed in the first example. Because once patients' recurrences have been detected, they belong to a specific treatment group based on the detection method, and the "treatment" is no longer time-dependent as in the first case. The last example shown in Chapter 6 is concerned with patients' quality of life and mental well-being during the trial. As a secondary outcome, the questionnaires were only filled out by patients at two distinct time points during the trial and an ANOVA-type model is used to make sensible inferences from the data.

Finally, Chapter 7 contains a summary and general discussion on the results of the thesis and discusses generalization and prospective future research.

REFERENCES

[1] Aoun SM, Grande G, Howting D, Deas K, Toye C, Troeung L, Stajduhar K, Ewing G (2015) The impact of the Career Support Needs Assessment Tool (CSNAT) in community palliative care using a stepped wedge cluster trial. PLoS One 10(4):e0123,012

[2] Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ (2015) Sample size calculation for a stepped wedge trial. Trials 16(1):354

[3] Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, et al (2015) Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials 16(1):1

[4] Britton B, McCarter K, Baker A, Wolfenden L, Wratten C, Bauer J, Beck A, McElduff P, Halpin S, Carter G (2015) Eating as Treatment (EAT) study protocol: a stepped-wedge, randomised controlled trial of a health behaviour change intervention provided by dietitians to improve nutrition in patients with head and neck cancer undergoing radiotherapy. BMJ

open 5(7):e008,921

[5] Brown BB, Young J, Smith DP, Kneebone AB, Brooks AJ, Xhilaga M, Dominello A, O'Connell DL, Haines M (2014) Clinician-led improvement in cancer care (CLICC)-testing a multifaceted implementation strategy to increase evidence-based prostate cancer care: phased randomised controlled trial-study protocol. Implementation Science 9(1):64

[6] Brown CA, Lilford RJ (2006) The stepped wedge trial design: a systematic review. BMC Med Res Methodol 6(1):1

[7] Bruinvels DJ, Stiggelbout AM, Kievit J, van Houwelingen HC, Habbema JD, van de Velde CJ (1994) Follow-up of patients with colorectal cancer. A meta-analysis. Ann Surg 219(2):174

[8] Campbell M, Donner A, Klar N (2007) Developments in cluster randomized trials and Statistics in Medicine. Stat Med 26(1):2–19

[9] Campbell MK, Elbourne DR, Altman DG (2004) CONSORT statement: extension to cluster randomised trials. BMJ 328(7441):702–708

[10] Fletcher SW (2011) Breast cancer screening: a 35-year perspective. Epidemiol Rev p mxr003

[11] Gambia Hepatitis Study Group and others (1987) The Gambia hepatitis intervention study. Cancer Res 47(21):5782–5787

[12] Girling AJ, Hemming K (2016) Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Stat Med 35(13):2149–2166, DOI 10.1002/sim.6850

[13] Grossmann I, Verberne C, de Bock G, Havenga K, Kema I, Klaase J, Renehan A, Wiggers T (2011) The role of high frequency dynamic threshold (HiDT) serum carcinoembryonic antigen (CEA) measurements in colorectal cancer surveillance: a (revisited) hypothesis paper. Cancers 3(2):2302–2315

[14] Hemming K, Taljaard M (2016) Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. J Clin Epidemiol 69:137–146

[15] Hemming K, Haines T, Chilton P, Girling A, Lilford R (2015) The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 350:h391

[16] Hemming K, Lilford R, Girling AJ (2015) Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Stat Med 34(2):181–196

[17] Highfield L, Rajan S, Valerio M, Walton G, Fernandez M, Bartholomew L (2015) A non-randomized controlled stepped wedge trial to evaluate the effectiveness of a multi-level mammography intervention in improving appointment adherence in underserved women. Implementation Science 10(1):143

[18] de Hoop E, van der Tweel I, van der Graaf R, Moons KG, van Delden JJ, Reitsma JB, Koffijberg H (2015) The need to balance merits and limitations from different disciplines when considering the stepped wedge clus-

ter randomized trial design. BMC Med Res Methodol 15(1):1

[19] Hughes JP, Granston TS, Heagerty PJ (2015) Current issues in the design and analysis of stepped wedge trials. Contemp Clin Trials 45:55–60

[20] Hussey MA, Hughes JP (2007) Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 28(2):182–191

[21] Jeffery M, Hickey BE, Hider PN, et al (2007) Follow-up strategies for patients treated for non-metastatic colorectal cancer. Cochrane Database Syst Rev 1(1)

[22] Juul JS, Bro F, Hornung N, Andersen BS, Laurberg S, Olesen F, Vedsted P (2016) Implementation of immunochemical faecal occult blood test in general practice: a study protocol using a cluster-randomised stepped-wedge design. BMC Cancer 16(1):445

[23] Kievit J (2000) Colorectal cancer follow-up: a reassessment of empirical evidence on effectiveness. Eur J Surg Oncol 26(4):322–328

[24] Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W (2012) Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. J Clin Epidemiol 65(12):1249–1252

[25] Lean M, Mann J, Hoek J, Elliot R, Schofield G (2008) Translational research

[26] Luckett T, Britton B, Clover K, Rankin N (2011) Evidence for interventions to improve psychological outcomes in people with head and neck cancer: a systematic review of the literature. Support Care Cancer 19(7):871–881

[27] Mdege ND, Man MS, Taylor CA, Torgerson DJ (2011) Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol 64(9):936–948

[28] Mdege ND, Man MS, Taylor CA, Torgerson DJ (2012) There are some circumstances where the stepped-wedge cluster randomized trial is preferable to the alternative: no randomized trial at all. response to the commentary by Kotz and colleagues. J Clin Epidemiol 65(12):1253

[29] Schwartz D, Lellouch J (2009) Explanatory and pragmatic attitudes in therapeutical trials. J Clin Epidemiol 62(5):499–505, DOI http://dx.doi.org/10.1016/j.jclinepi.2009.01.012

[30] Tjandra JJ, Chan MK (2007) Follow-up after curative resection of colorectal cancer: a meta-analysis. Diseases of the colon & rectum 50(11):1783–1799

[31] Toftegaard BS, Bro F, Vedsted P (2014) A geographical cluster randomised stepped wedge study of continuing medical education and cancer diagnosis in general practice. Implementation Science 9(1):159

[32] Verberne C, Zhan Z, van den Heuvel E, Grossmann I, Doornbos P, Havenga K, Manusama E, Klaase J, van der Mijle H, Lamme B, et al (2015) Intensified follow-up in colorectal cancer patients using frequent Carcino-Embryonic Antigen (CEA) measurements and CEA-triggered imaging: Results of the randomized 'CEAwatch'

trial. Eur J Surg Oncol 41(9):1188–1196

[33] Verberne CJ, Nijboer CH, de Bock GH, Grossmann I, Wiggers T, Havenga K (2012) Evaluation of the use of decision-support software in carcino-embryonic antigen (CEA)-based follow-up of patients with colorectal cancer. BMC Med Inform Decis Mak 12(1):14, DOI 10.1186/1472-6947-12-14

[34] Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S (2013) Stepped wedge designs could reduce the required sample size in cluster randomized trials. J Clin Epidemiol 66(7):752–758

[35] World Health Organization (1983) Prevention of liver cancer: report of a WHO meeting [held in Geneva from 30 January to 4 February 1983]

[36] Yu X, Klesges LM, Smeltzer MP, Osarogiagbon RU (2015) Measuring improvement in populations: implementing and evaluating successful change in lung cancer care. Translational lung cancer research 4(4):373

2

STRENGTHS AND WEAKNESSES OF A STEPPED WEDGE CLUSTER RANDOMIZED DESIGN: ITS APPLICATION IN A COLORECTAL CANCER FOLLOW-UP STUDY

Z. Zhan
E. R. van den Heuvel
P. M. Doornbos
H. Burger
C. J. Verberne
T. Wiggers
G. H. de Bock

This chapter has been published in Journal of clinical epidemiology, Volume **67**, Issue 4, (2014) [26].

2

ABSTRACT

Objectives: To determine the advantages and disadvantages of a stepped wedge design for a specific clinical application.

Study design and settings: The clinical application was a pragmatic cluster randomized surgical trial intending to find an increased percentage of curable recurrences in patients in follow-up after colorectal cancer. Advantages and disadvantages of the stepped wedge design were evaluated, and for this application, new advantages and disadvantages were presented.

Results: A main advantage of the stepped wedge design was that the intervention rolls out to all participants, motivating patients and doctors, and a large number of patients who were included in this study. The stepped wedge design increased the complexity of the data analysis, and there were concerns regarding the informed consent procedure. The repeated measurements may bring burden to patients in terms of quality of life, satisfaction, and costs.

Conclusions: The stepped wedge design is a strong alternative for pragmatic cluster randomized trials. The known advantages hold, whereas most of the disadvantages were not applicable to this application. The main advantage was that we were able to include a large number of patients. Main disadvantages were that the informed consent procedure can be problematic and that the analysis of the data can be complex.

2.1. INTRODUCTION

Back in 1967, the concept of a pragmatic trial was proposed by Schwartz and Lellouch.[19] A pragmatic trial is designed to evaluate the effectiveness of interventions in real-life routine practice conditions [16]. Its aim is to answer the question "Does the intervention work when used in real-life practice condition?" Thus, in a pragmatic trial, there are no or minimal exclusion criteria required. It also implies that pragmatic trials are normally used when there is a priori knowledge of the efficacy of the intervention under study. In addition, a pragmatic trial is often concerned with complex interventions, for example, routine screening of the disease rather than a pharmaceutical intervention, and it is typically compared with care as usual.

To answer the question whether the intervention works when used in real-life practice, it is very common to apply a cluster randomized design. One of the reasons for choosing a cluster design is the concern of contamination.[23] Cluster randomization may reduce the risk that the intervention under study is unintentionally mixed up with care as usual, the intervention of the control group.[1, 18, 22] Another reason is that the intervention can be performed more easily in clusters as a large number of participants can make it impractical to introduce a new treatment on an individual level, for example, when medical resources are low or when there are costly expenses.[6]

The stepped wedge design is a unique design suitable to answer the question whether the intervention works when used in real-life practice.[3, 10] This design allows for a controlled stepwise introduction of an intervention to a population.[3, 10] Although not per definition, the stepped wedge design is a design that is mostly performed as a cluster design.[15] In this design, participants start in the control group, and at predefined time points, a cluster of participants are switched to the intervention

group in a random order (known as “steps”). From the moment of switching until the end of the study, they will stay in the intervention group.[10]

When the stepped wedge design is compared with other designs, there are several advantages and disadvantages of choosing such a design. The aim of the present article was to determine the advantages and disadvantages of a stepped wedge cluster randomized design for a specific clinical application. An overview of the literature regarding the advantages and disadvantages of the stepped wedge design will be given. The clinical application is the colorectal cancer (CRC) follow-up study (CEAwatch, Netherlands Trial Register 2182). Based on a point-by-point evaluation, it was analyzed how these advantages and disadvantages apply to our specific clinical trial. Furthermore, some new trial-specific advantages and disadvantages of the stepped wedge design in this application, which were not mentioned in the literature, will be added.

2.2. METHODS/DESIGN

2.2.1. ADVANTAGES AND DISADVANTAGES OF STEPPED WEDGE DESIGNS

Under the circumstance when there is prior evidence that the intervention under study will do more good than harm, rather than clinical equipoise, it is considered not ethical to withhold or withdraw an intervention from participants.[3] As the stepped wedge design provides unidirectional sequential rollout of the intervention, all participants will get the intervention during the study. Additionally, the stepped wedge design can be a good option in trials in which it is not possible to introduce the intervention to all participants at once because of logistic, financial, or practical reasons as the design introduces the intervention over multiple moments.[3] The stepped wedge design is considered to be a strong design to evaluate effects on a population level.[9] It is favored over some other trial designs because it provides an opportunity to measure possi-

ble effects of the time of the intervention and to investigate the effects of underlying temporal changes.[3] The stepped wedge design is more efficient than others as it may reduce the required number of clusters compared with other classic cluster designs.[10, 25] Although between-cluster variation affects the statistical power in a parallel clustered randomized design, the power appears to be relatively insensitive to between-cluster variation in the stepped wedge design.[10] Stepped wedge design requires fewer clusters because the power of the design is mainly determined by within-cluster variations.[10]

One of the drawbacks of the stepped wedge design is that it takes a longer time to perform.[13] Because of the nature of a stepwise introduction of the intervention, the trial duration of a stepped wedge design will be the duration of a classic cluster randomized trial multiplied by the number of steps. Especially for clusters that started later have to wait longer depending on the duration of each step, it may cause them to switch into interventions or dropping out. This will then increase the risk of attrition. In addition, the repeated measurements of the stepped wedge design put a heavy burden on patients, caregivers, and researchers.[13] Another concern is that the stepped wedge design may increase the risk of contamination in a cluster, especially when the intervention is believed to be superior to control.[13] It is also very hard to use blinding because both patients and assessors are aware of the step switch.[3] From a statistical perspective, there are also some disadvantages of using the stepped wedge design. Mentioned by Hussey [10], a delay in the treatment effect reduces the power of the design. Moreover, the analysis of the stepped wedge design is more complex.[15]

For a summary of the literature on the advantages and disadvantages of the stepped wedge design, see Table 2.2, first column. Based on a point-by-point evaluation, it was analyzed how these advantages and disadvantages applied to the specific clinical trial CEAwatch (Netherlands

Trial Register 2182). Furthermore, some application-specific advantages and disadvantages of the stepped wedge design, which has not been mentioned in literature, were added.

2.2.2. CLINICAL APPLICATION TO THE CRC FOLLOW-UP (CEAWATCH)

The tumor marker carcinoembryonic antigen (CEA) has long been known to be important in signaling recurrent disease in CRC.[20] Intensive follow-up schedules including CEA measurements are correlated with better survival rate than schedules not using CEA measurements [11], and serial measurements of CEA are recommended for use in CRC follow-up.[4, 14] Other studies also confirmed a reduction of mortality rate and an improvement in curative reoperation rate with intensive surveillance.[21] In a phase 2 trial, monthly CEA measurements were done with a threshold of two consecutive rises of more than 10%.[7] The trial showed both high sensitivity and specificity for detection of recurrences using serial CEA rises rather than absolute values. Given this evidence, in CEAwatch, a new intensified follow-up scheme including frequent CEA measurements and CEA-triggered imaging in detecting recurrent disease with curative possibilities in CRC patients was compared with care as usual.

PATIENTS

Patients with American Joint Committee on Cancer stage I, II, and III CRC after R0 resection, who were surgically operated, were eligible. Patients who received adjuvant chemotherapy were eligible after termination of adjuvant therapy. Patients who were not medically fit for metastasectomy, patients diagnosed with other malignancies (except skin basocellular carcinoma), and patients with metachronous metastases at the start of the study were excluded.

THE FOLLOW-UP CARE AS USUAL

The control or "care-as-usual" follow-up consisted of follow-up as recommended in the national guideline in the Netherlands (www.tinyurl.com/coloncarcinoma) including an outpatient clinic visit every 6 months for the first 3 years and an annual visit in years 4 and 5. Liver ultrasound and chest x-ray were recommended at each clinic visit. CEA was measured every 3–6 months in the first 3 years and each year in the last 2 years. No monitoring of compliance with this recommendation was provided.

THE CEAWATCH FOLLOW-UP

The intensified follow-up protocol adhered to bimonthly CEA measurements and yearly imaging in the first 3 years and trimonthly CEA measurements in the fourth and fifth years of follow-up (Fig. 2.1). Outpatient clinic visits with imaging of chest and abdomen were performed annually in the first 3 years. The threshold value used was a 20% rise compared with the latest CEA value, followed by a threshold of any rise respect to the last measurement after 1 month. In case of two consecutive rises in CEA, a computed tomographic (CT) scan of chest and abdomen was advised for localization of potential metastatic disease. The coordination of this process was supported by an automatic computer system.[20, 24] Doctors were given an alert when a CT scan was indicated because of a consecutive rise in CEA or when patients forget to go for a CEA assessment. CEA values were communicated to the patients by an automatically generated letter, including a laboratory form for the next CEA measurement.

STUDY DESIGN

The hospitals were randomly grouped into five clusters that were changed from the usual follow-up schedule to the intensive follow-up schedule at different time points. Cluster crossover from the control schedule to the intervention schedule occurred in one direction only and once every 3 months (Table 2.1). Randomization of the crossover moments of

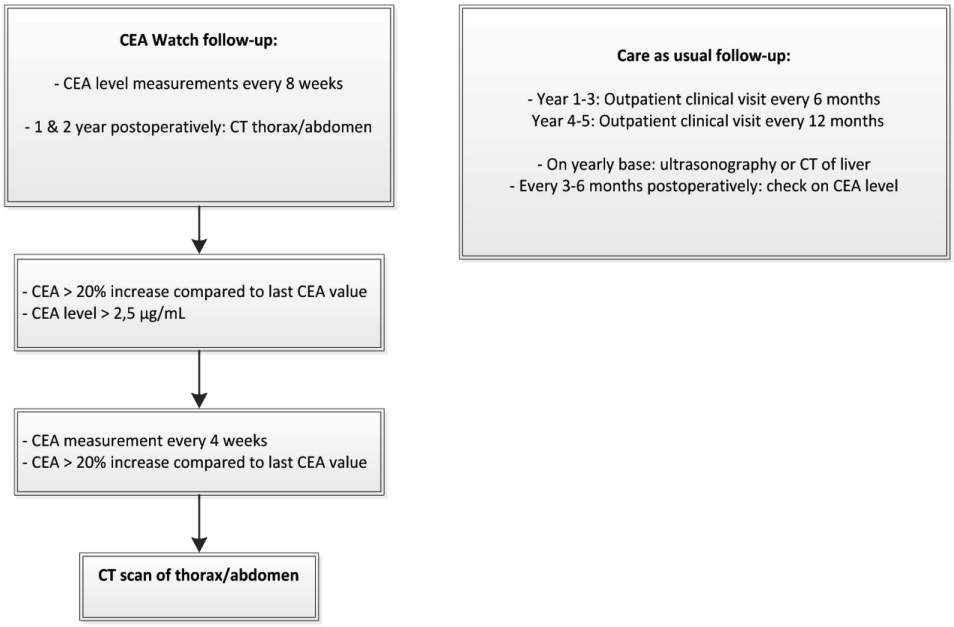

Figure 2.1 | CEA-Watch follow-up and the care-as-usual follow-up. *Local differences and adjustment by individual hospitals allowed.

the clusters was performed independently by Trial Coordination Center Groningen (`www.tcc.umcg.nl`). CEAwatch was approved by the Medical Ethics Committee of the University Medical Center Groningen and the local ethics committees of all participating centers. For an overview of the procedures, see Fig. 2.2.

MAIN OUTCOME

The primary outcome measures were the proportion of resectable recurrences among all recurrences and the time to and probability of detection of recurrent disease in the intervention protocol compared with the control protocol.

DATA COLLECTION

In the participating hospitals, the eligible patients were identified using the diagnosis or operation code(s). At the end of the study, this search

Table 2.1 | Progression of control (0) and intervention group (1) over time periods (t) in CEA-Watch study

Cluster of hospitals	October 2010–January 2011	January 2011–April 2011	April 2011–July 2011	July 2011–October 2011	October 2011–January 2012	January 2012–October 2012	Number of Patients
A	0	1	1	1	1	1	721
B	0	0	1	1	1	1	456
C	0	0	0	1	1	1	613
D	0	0	0	0	1	1	630
E	0	0	0	0	0	1	803
Number of participants	2,498	2,484	2,503	2,409	2,255	1,946	3,223

was validated against the database of the Dutch Comprehensive Cancer Center. In this database, all newly diagnosed malignancies are registered based on the automated pathologic archive. After all eligible patients were identified, patient and tumor characteristics were exported from the Dutch Surgical Colorectal Audit (DSCA) into a password-protected database. DSCA is an obligatory national data bank that gathers all relevant information on surgically treated CRC patients, allowing a valid registration of all CRC patients in the Netherlands, without any missing baseline characteristics (`www.clinicalaudit.nl/dsca`). Per hospital, there was one study coordinator. The study coordinators were uniformly trained to identify new eligible patients, inform patients about the study, and collect the follow-up data. The study coordinators were continuously monitored by one of the investigators.

POWER CALCULATION

The expected percentage of resectable recurrences was 10% in the control protocol and 25% in the intensified protocol.[2, 17] Given a significance level of 5% and a power of 80%, 115 patients with recurrent disease in both groups were needed. Given an expected recurrence rate of 25% [12],

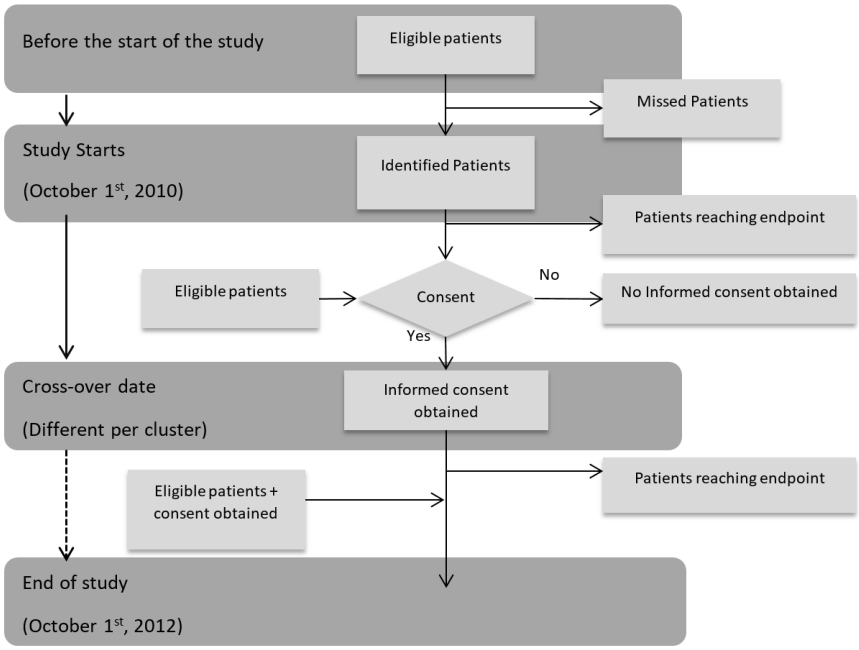

Figure 2.2 | Study procedures in CEA-Watch.

460 patients per group were needed. Given the cluster randomization, we assumed a correlation of 0.1 between hospitals, yielding a correction factor of 1.71.[8] Therefore, a minimum of about 800 patients per group was needed.

DATA ANALYSIS

To compare the effect of the intensified follow-up with the control follow-up protocol regarding the proportion of resectable recurrences, a conditional logistic regression analysis, with hospital as the stratification variable, was performed. Cox proportional hazards model formed the basis of the analysis of the time-to-event data (recurrence or curable recurrence). Hereby, the follow-up protocols were used as a time-dependent variable because the switch time in follow-up was dynamic for patients. The time from the operation until the participation in the study created

left truncated data, and to correct for this, a delayed entry variable was implemented. Again, the analysis was stratified by hospital.

CURRENT STATUS OF CEAWATCH

Inclusion of patients (n = 3223) started from October 2010 and ended in July 2012. Every 3 months, there was a switch from the care-as-usual follow-up to CEAwatch follow-up, which successfully took place. The first switch was in January 2011. The results from this study will be published separately.

2.3. RESULTS

2.3.1. ADVANTAGES

The logistic difficulties of implementing the intervention everywhere at once was one of the considerations to choose for a stepped wedge design for the CEAwatch study (Table 2.2).[3] To start the study in 11 hospitals, approvals from 11 local administration institutes were needed. In each participating hospital, the eligible patients had to be identified, patient and tumor characteristics had to be extracted, and study coordinators had to be trained to be able to identify new eligible patients, inform patients about the study, and collect the follow-up data before the study could start. Besides that, the automatic computer system that was used to support the implementation of the intervention under study had to be adapted to the local hospital system before it could be implemented in the hospitals. Thus, it was not feasible to implement the intervention to different hospitals simultaneously.

The stepped wedge design is considered more ethical than other (classic) cluster randomized controlled designs when the intervention is believed to do more good than harm.[3] In our case, there was enough evidence to support the follow-up of patients with CRC with frequent

Table 2.2 | Advantages and disadvantages of stepped wedge design and their application to CEAwatch

General	Application of general advantages/disadvantages to CEAwatch
Advantages	
Good alternative when interventions cannot be implemented to all clusters simultaneously because of practical, logistic, or financial constraints.[3]	CEAwatch involved 11 hospitals all around the country, it was very hard to implement the intervention simultaneously, and the specialized software used in this trial also took time to be adjusted hospital by hospital. The stepped wedge design provides an opportunity to prepare for the implementation in the control period during the trial.
If there is a prior belief that the intervention will do more good than harm (rather than clinical equipoise), it is considered not ethical to withhold/withdraw intervention from participants. The stepped-wedge design provides sequential rollout of the intervention for all participants.[3]	The CEA measurements and the intensive follow-up protocol were proven effective from individual-level trials. This point of advantage was motivation for patients and doctors to participate. This helped to increase the size of the sample and consequentially increased the power of the trial.
Provide opportunity to measure possible effects of time of intervention and to investigate the effects of underlying temporal changes.[3]	The end points were time to event and events, which need a certain time period before they are observed. In combination with the relative short period (3 months) between switches, the underlying temporal changes could not be appropriately investigated. Because of the longitudinal setup of the CEAwatch, this point is not applicable for the study.
Reduces the number of clusters.[10]	CEAwatch intended to include as many hospitals as possible and has no issue regarding the number of clusters. Thus, the study did not make use of this benefit.
It is more efficient than other cluster randomized controlled trial design.[25]	Thus far, this claim was proven under circumstances of very simple settings of a stepped wedge design. It is not sure whether a stepped wedge design will be a more efficient design than other designs in our more complex setting (left truncation, dynamic, or time-varying intervention and stratification in the Cox proportional hazards approach).

General	Application of general advantages/disadvantages to CEAwatch
Disadvantages	
Longer trial durations and increased risk of attrition.[13]	Because of the required longitudinal setup for the CEAwatch trial (multiple visits), using the stepped wedge design does not extend the trial duration compared with other designs. The follow-up for eligible patients is 5 years. It is essential to have comparable trial duration (eg, 3–5 years) to investigate the effectiveness of the intensive follow-up routine no matter what kind of design is being used.
Repeated measurements put a heavy burden on patients, caregivers, and researchers.[13]	In a cluster randomized trial, not all patients have to go through the intensified follow-up. This would be beneficial when the intensified approach would not be as effective as it was anticipated. Whether patients also view the CEAwatch intensified follow-up as a higher burden is very critical, thus we wanted to measure this with quality of life and cost-effectiveness studies.
Increased risk of contamination.[13]	The risk of contamination was limited because of the implementation of tailor-made software that would support the physicians in the intensified CEA follow-up.
Delay in treatment effect reduces the power of the design.[10]	This is true for CEAwatch, but considering the benefits from a substantial larger sample size, the power reduction from delay in treatment does not have big influence.
Lack of blinding	Information bias can also be considered as part of responses of treatment of patients and physicians for pragmatic trials such CEAwatch
Analysis of the design is complex.[15]	The CEAwatch has a rather complex design, we consider this point as the main disadvantage of stepped wedge design
CEAwatch specific	
Advantage	Recruitment of participants was much easier during the CEAwatch. This allows hospitals to enter the control period with same criteria of eligible patients and include new patients during the study period.
Disadvantage	Asking informed consent from all patients at baseline was not approved by the Ethics Committee of our hospital. It is considered not acceptable for patients.

CEA measurements and CEA-triggered imaging in detecting recurrent disease with curative possibilities. Because of the preferences of surgeons for the intervention under study, this ethical advantage was an important motivation for doctors to participate in the trial. As we had no published evidence that the repeated CEA measures in the intervention under study was not a burden to the patient associated with an increase of costs, data on secondary outcomes such as patient satisfaction, quality of life, and costs were collected.

Another advantage of the stepped wedge design is that this design provides an opportunity to measure possible effects of time of the intervention and investigate the effects of underlying temporal changes because of its longitudinal settings.[3] The CEAwatch study could not really benefit from this point because the longitudinal setting was needed to obtain or collect the events. Besides this, the periods before the switches were relatively short (only 3 months), making it less attractive to model time trends in the analysis of the events. Furthermore, the inclusion of patients was dynamic, complicating such a temporal analysis.

Another general advantage of the stepped wedge design is that it reduces the required number of clusters as the design is relatively insensitive to variations of the intercluster correlation.[10] This might be beneficial for trials that have limited resources and cannot include enough clusters, but for the CEAwatch study, the number of clusters was sufficient and there was no need to include as many hospitals as possible. Thus, this advantage was not one of the considerations to choose for the stepped wedge design. It is claimed that a stepped wedge design is more efficient than other cluster randomized controlled designs.[25] Thus far, this claim was proven under circumstances of very simple settings of a stepped wedge design. It is not sure whether a stepped wedge design was a more efficient design than other designs in our more complex application, which required left truncation, a dynamic or time-varying intervention variable,

and stratification in the Cox proportional hazards approach.

An advantage of the stepped wedge design not mentioned in literature, but very important in our study, is that by the use of the stepped wedge design, we were able to include a large number of patients. In the CEAwatch study, eligible patients were identified before the start of the study, and new patients were included during the time of the study. An advantage of this approach was that patient selection was less vulnerable to selection bias. A second advantage of this approach was that the group of participants consisted of patients who were already in follow-up on the date of the start of the study and those who became eligible during the study period.

2.3.2. DISADVANTAGES

One of the disadvantages of the stepped wedge design is that it takes longer than the more traditional designs.[13] However, because of the longitudinal setup of the CEAwatch study, using a stepped wedge design did not extend the trial duration compared with other designs. As the follow-up for eligible patients was in principle 5 years, it was essential to have comparable trial duration (eg, 3–5 years) to investigate the effectiveness of the intensive follow-up routine no matter what kind of design would have been used. As a consequence, in this case, other designs would not have shortened the trial duration substantially.

Another drawback of the stepped wedge design is the heavy burden on patients, caregivers, and researchers caused by the necessary repeated measurements.[13] As these repeated measurements could not be avoided because of the nature of the intervention in the CEAwatch study, other designs would also have had this problem. On the other hand, in a cluster randomized trial, not all patients have to go through the intensified approach. This could be beneficial when the intensified approach would not be as effective as it was hypothesized. Whether patients also viewed

the CEAwatch intensified follow-up as a burden is of course critical, it was decided to investigate this in the study as secondary outcomes.

Another concern mentioned by Kotz et al. [13] is the increased risk of contamination and attrition. The contamination in a cluster in the CEAwatch study was limited to a minimum because of the automated software system that was used to trigger follow-up schedules. Thus, it would have been very hard that certain patients would still be scheduled under the care-as-usual when the hospital would have changed to the intensified intervention. The attrition problem in the study was mainly due to the long trial duration, which would most likely also have occurred in other types of designs.

When designing a stepped wedge and estimating its sample size, it is suggested by Hussey and Hughes [10] that researchers should take into account the delay in treatment effect as the effect of such delay is a reduction of the power of the design. However, given the nature of the outcome, the delay in treatment effect was considered to be not a problem for the CEAwatch study. It was somewhat compensated with the inclusion of patients who had surgery before the start of the study.

Although it is true that using blinding in the CEAwatch is impossible, it is not typical for a stepped wedge design. In addition, the effects might be not as strong as it is claimed to be.[3] As a pragmatic trial, CEAwatch was interested in studying the responses of patients in a real-life situation. As a consequence, the awareness of the intervention could be accepted as part of the responses to treatment.

Furthermore, it is mentioned that the analysis of the stepped wedge design is complex.[15] This was considered a main disadvantage of the design in CEAwatch. The complexity comes from different sources. One source is the issue with the delayed entry of patients into the study, and another source is the dynamic nature of the inclusion of patients and the switch moments that require a time-varying intervention variable into

the survival analysis. The hospitals were addressed by stratification in the analyses, but they may also be considered as random, which would be typical in the more classical cluster randomized trials. This approach may complicate the analysis. Although we believe that the analysis might be reasonable, more research on the statistical analysis is required to verify if the estimate of the intervention effect is not biased.

Another disadvantage not mentioned in literature is related to the timing of the informed consent procedure. When the intention was to ask informed consent from all patients at baseline, this could not be realized. The reason was that the Medical Ethics Committee of our hospital did not consider this as acceptable for patients. Therefore, patients were asked for informed consent before the switch from the control to the intervention period. The patients who entered the study after the switch were asked for informed consent before surgery. As it was impossible to ask all patients for informed consent at the outpatient ward in the few weeks before switching follow-up, letters were generated for this purpose. Consequently, patients who did not respond to the letter were not included in the intervention period, making the intervention group smaller than expected. Patients who do not response to the letter or exit the study during the control period (eg, because of patient death or having a recurrence) were not asked for informed consent. However, their data could still be used. This was possible as these patients did not experience any changes in follow-up and had a guaranteed anonymity (according to the Dutch law) by the assignment of unique patient numbers and a password-protected database.

2.4. DISCUSSION

Not only the stepped wedge design helps with the implementation difficulty, it is also considered more ethical because there is enough evidence

to support the efficacy of the intervention. Another advantage is that the rollout setting for all participants of the stepped wedge design motivates not only the patients but also the doctors to participate in this study. Recruitment of participants was therefore much easier in CEAwatch study. This allows hospitals to enter the control period with the same criteria of eligible patients and include new patients during the study period. This advantage has not been emphasized in literature yet. Furthermore, the sequential introduction of the intervention was a real benefit to the CEAwatch study. It would have been almost impossible to select another trial design. Other generally accepted advantages such as opportunity of time effect investigation and reduction to the number of clusters did not have the expected benefit for the study. On the other hand, the application of the stepped wedge design to CEAwatch increased the complexity in data analysis and the repeated measurements may bring additional burden to patients in terms of quality of life, satisfaction, and costs. In addition, there are concerns regarding the procedure of informed consents. This trial-specific disadvantage of steped wedge design is new to those general ones.

Because the analysis of the study is still in progress, whether using a stepped wedge design provides unbiased estimation of the treatment effect remains to be further investigated. To the extent that missing data are negligible, we believe with proper analysis method that the estimation should be adequately unbiased.

We were able to include a large number of patients. In many surgical trials, the inclusion of patients is one of the key problems. This is also a good solution to the challenge of recruitment difficulties mentioned in surgical studies. This challenge is mainly due to the strong preferences of patients and surgeons for one intervention and the organization of the inclusion and randomization.[5]

Because the original sample size calculation did not take into account

the design effect of the stepped wedge design, a more correct sample size calculation by Hussey and Hughes [10] was performed retrospectively. The minimal number of patients per cluster per time interval was determined at 187. We used the same input information that was described earlier.

The analysis of the benefits and drawbacks brought by the stepped wedge design indicates that it is a strong alternative for pragmatic cluster randomized trials such as the CEAwatch. The general advantages of the stepped wedge design still holds compared with other controlled trial design, whereas most of the general concerns regarding the stepped wedge design bring no disadvantages to the CEAwatch study. However, the stepped wedge design makes the analysis of the trial rather complex and whether repeated measurements bring burden to patients needs further investigation. One advantage that has not been mentioned in literature before is that the stepped wedge design contributes to larger sample size because of not only the ethical advantage of the design but also the rollout setting, which provides strong motivation for doctors. This allows hospitals to enter the control period with same criteria of eligible patients and include new patients during the study period. Meanwhile, difficulty in the informed consents was found as a disadvantage specifically in our clinical application.

REFERENCES

[1] Altman DG (1990) Practical statistics for medical research. CRC press

[2] Bentrem DJ, DeMatteo RP, Blumgart LH (2005) Surgical therapy for metastatic disease to the liver. Annu Rev Med 56:139–156

[3] Brown CA, Lilford RJ (2006) The stepped wedge trial design: a systematic review. BMC Med Res Methodol 6(1):1

[4] Duffy M, van Dalen A, Haglund C, Hansson L, Holinski-Feder E, Klapdor R, Lamerz R, Peltomaki P, Sturgeon C, Topolcan O (2007) Tumour markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines for clinical use. Eur J Can-

cer 43(9):1348–1360

[5] Ergina PL, Cook JA, Blazeby JM, Boutron I, Clavien PA, Reeves BC, Seiler CM, Collaboration B, et al (2009) Challenges in evaluating surgical innovation. The Lancet 374(9695):1097–1104

[6] Gambia Hepatitis Study Group and others (1987) The Gambia hepatitis intervention study. Cancer Res 47(21):5782–5787

[7] Grossmann I, Verberne C, de Bock G, Havenga K, Kema I, Klaase J, Renehan A, Wiggers T (2011) The role of high frequency dynamic threshold (HiDT) serum carcinoembryonic antigen (CEA) measurements in colorectal cancer surveillance: a (revisited) hypothesis paper. Cancers 3(2):2302–2315

[8] van Houwelingen J (1998) Roaming through methodology. III. Randomization at the level of the physicians. Ned Tijdschr Geneeskd 142(29):1662–1665

[9] Hughes J, Goldenberg RL, Wilfert CM, Valentine M, Mwinga KG, Guay LA, Mmiro F, Stringer JS (2003) Design of the HIV prevention trials network (HPTN) protocol 054: a cluster randomized crossover trial to evaluate combined access to nevirapine in developing countries. Tech. Rep. Working Paper 195, UW Biostatistics Working Paper Series.

[10] Hussey MA, Hughes JP (2007) Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 28(2):182–191

[11] Jeffery M, Hickey BE, Hider PN, et al (2007) Follow-up strategies for patients treated for non-metastatic colorectal cancer. Cochrane Database Syst Rev 1(1)

[12] Kobayashi H, Mochizuki H, Sugihara K, Morita T, Kotake K, Teramoto T, Kameoka S, Saito Y, Takahashi K, Hase K, et al (2007) Characteristics of recurrence and surveillance tools after curative resection for colorectal cancer: a multicenter study. Surgery 141(1):67–75

[13] Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W (2012) Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. J Clin Epidemiol 65(12):1249–1252

[14] Locker GY, Hamilton S, Harris J, Jessup JM, Kemeny N, Macdonald JS, Somerfield MR, Hayes DF, Bast Jr RC (2006) ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. J Clin Oncol 24(33):5313–5327

[15] Mdege ND, Man MS, Taylor CA, Torgerson DJ (2011) Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol 64(9):936–948

[16] Patsopoulos NA (2011) A pragmatic view on pragmatic trials. Dialogues in clinical neuroscience 13(2):217

[17] Pfannschmidt J, Dienemann H, Hoffmann H (2007) Surgical resection of pulmonary metastases from colorectal cancer: a systematic review of published series. The Annals of thoracic surgery 84(1):324–338

[18] Pocock SJ (2013) Clinical trials: a practical approach, John Wiley & Sons, chap Methods of Randomization

[19] Schwartz D, Lellouch J (2009) Explanatory and pragmatic attitudes in therapeutical trials. J Clin Epidemiol 62(5):499–505, DOI http://dx.doi.org/10.1016/j.jclinepi.2009.01.012

[20] Staab HJ, Anderer FA, Stumpf E, Fischer R (1978) Slope analysis of the postoperative CEA time course and its possible application as an aid in diagnosis of disease progression in gastrointestinal cancer. The American Journal of Surgery 136(3):322–327

[21] Tjandra JJ, Chan MK (2007) Follow-up after curative resection of colorectal cancer: a meta-analysis. Diseases of the colon & rectum 50(11):1783–1799

[22] Torgerson DJ (2001) Contamination in trials: is cluster randomisation the answer? BMJ 322(7282):355

[23] Treweek S, Zwarenstein M (2009) Making trials matter: pragmatic and explanatory trials and the problem of applicability. Trials 10(1):37

[24] Verberne CJ, Nijboer CH, de Bock GH, Grossmann I, Wiggers T, Havenga K (2012) Evaluation of the use of decision-support software in carcino-embryonic antigen (CEA)-based follow-up of patients with colorectal cancer. BMC Med Inform Decis Mak 12(1):14, DOI 10.1186/1472-6947-12-14

[25] Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S (2013) Stepped wedge designs could reduce the required sample size in cluster randomized trials. J Clin Epidemiol 66(7):752–758

[26] Zhan Z, van den Heuvel ER, Doornbos PM, Burger H, Verberne CJ, Wiggers T, de Bock GH (2014) Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. J Clin Epidemiol 67(4):454–461

3

STATISTICAL ANALYSIS OF THE STEPPED WEDGE DESIGN: A PRATICAL NOTE WITH SIMULATIONS

Z. Zhan
E. R. van den Heuvel
G. H. de Bock

This chapter is submitted to American Journal of Epidemiology

ABSTRACT

A stepped wedge design is a randomized controlled trial design with a phased introduction of the intervention at different moments of the trial. It has attracted a lot of interests from medical researchers and epidemiologists and much debate has been focusing on the practical aspect of the design while the statistical analysis was less addressed. The objective of the presented study is to evaluate statistical methods for analysing binary outcome data arising from a stepped wedge clustered randomized trial in a systematic and expository manner. We included statistical methods that are not commonly considered in the stepped wedge design literature and highlight some of the limitations of the current commonly used methods. Specifically, we considered an aggregate-data meta-analysis approach when no period effect exists, a marginal model with generalized estimating equations at a cluster level, Hussey and Hughes variance components model at both individual level and cluster level, and a subject specific growth model at individual level. Simulations were conducted to compare the performances of these methods under varying assumptions about the period effects and period-treatment interaction effects. Simulation results showed that the marginal model and the meta-analysis approach were both valid choices as aggregated-level data analysis approach, but the former one can also be used when a period effect is present. Furthermore, the linear mixed model of Hussey and Hughes provided biased estimates of the treatment effect when a period-treatment interaction was ignored. Even when a period-treatment interaction would be taken into account, the model still did not have the correct interpretation due to parametrization issues, possibly leading to incorrect inferences in practice. A subject-specific growth model that took time period as continuous variable had a straightforward parametrization and interpretation but was prone to misspecifications of the period effect.

3.1. INTRODUCTION

A stepped wedge design is a randomized controlled clinical trial design which utilizes a randomized sequential roll-out of the intervention.[12] At the beginning of the trial, participants will start with the control treatment. Switching the treatment to the new intervention will take place at predetermined switching moments. At the end of the trial, the treatment arm consists of the new intervention only. Though not per definition, the stepped wedge design is most often randomized at cluster level. Therefore, the clustered randomized stepped wedge design will be the focus of the presented paper. Advantages of the stepped wedge design, including logistical flexibilities, efficiencies in terms of power and sample size compared to traditional (clustered) parallel group designs, and the ethical advantages in longitudinal and open cohort studies, have been recognized.[1–3] Therefore, stepped wedge designs have been increasingly adopted by medical researchers. As such, much debates have been focusing on the practical aspects of the design.[3, 7, 15, 21, 23, 32]

However, the statistical methodology for the analysis of the stepped wedge design is still "in its infancy".[11] In terms of data analysis methods, stepped wedge designs are quite unique. Unlike parallel group designs, randomization units in the stepped wedge design are no longer being allocated to distinct treatment arms for comparison. Stepped wedge designs are also different from crossover designs since the switching is in one direction only. Due to the unique character of the stepped wedge design, it is unclear and sometimes confusing which statistical methods should be used in a stepped wedge design.[8] The most frequently applied statistical methods are developed by Hussey and Hughes.[20] However, their linear mixed model assumes a constant treatment effect across clusters and over periods and they may not be applicable for all studies. A recent study found that the estimated treatment effects had up to 50%

biases when models are misspecified in the presence of heterogeneous treatment effects at different clusters.[25]

The literature is sparse on the analysis of stepped wedge designs when treatment-period interactions would be present. Additionally, outcomes other than normally distributed variables have hardly been addressed. Alternative analysis methods, other than Hussey and Hughes [20](and its extension by Hemming and Girling [10]), have not been proposed yet either. This indicates the need to study current methods under more realistic settings and to expand the tool set for analysis of stepped wedge design. Therefore, we will use simulation studies to illustrate some of the commonly proposed data analysis methods for the stepped wedge design and examine the validity of these methods.

3.2. METHODS

Data analysis of stepped wedge designs can be broken down into two distinct categories. Either the data can be analysed at an aggregated level by taking summary measures at cluster-period combination; or the data can be analysed at an individual level. Both approaches have to address possible confounding issues of period and treatment effects, since in the stepped wedge design, period is associated with both the outcome and the treatment. Several approaches within the two categories will be discussed in more detail.

3.2.1. AGGREGATED-DATA ANALYSIS

In clustered randomized trials, outcomes may be summarized into one measure for each cluster across the whole period [4] but in clustered randomized stepped wedge design studies, a cluster-period measure is needed to be able to deal with treatment and period effects. Let Y_{ijk} be the result of participant k observed in cluster i at period j. The aggre-

gate measure is then $M_{ij} \equiv M(Y_{ij1}, \cdots, Y_{ijk_{ij}})$ with M_{ij} for instance the average or median.

WITHOUT PERIOD EFFECTS

In case there is no period effect, the aggregate measure can be further summarized over the periods that belong to the same treatment. This results into a pair (M_i^C, M_i^T) for the control and treatment at cluster i. In this way, each cluster will contribute to the overall effect size. For instance, for continuous outcomes Y_{ijk}, the effect size can be the difference $M_i^C - M_i^T$ or the ratio M_i^T / M_i^C, but for binary outcomes Y_{ijk} the odds ratio $M_i^T(1 - M_i^C)/M_i^C(1 - M_i^T)$, with M_i^X the average for treatment X, can be used. Consequently, the whole trial can be viewed as a meta-analysis study and any appropriate method for synthesizing effect sizes can be applied to this situation (see for example [9]). However, such approach does require an appropriate estimator for the standard error of the selected effect size. If it is calculated from the standard errors on M_i^C and M_i^T, it is important to mention that the standard errors for M_i^C and M_i^T will not be the same, as the measures will be calculated from different numbers of observations. The numbers of periods before the switch are typically different from the numbers of periods after the switch. Alternatively, meta-analysis methods can also be directly applied to the pair (M_i^C, M_i^T) considering a joint distribution. Such joint model is typically preferred for binary outcomes Y_{ijk}. The measure M_i^X will be taken as the number of events for treatment X and the pair (M_i^C, M_i^T) will be considered independently binomially distributed conditionally on cluster i.[16, 17] A pooled odds ratio is then estimated from this generalized linear mixed model, which would indicate the treatment effect.

Hussey and Hughes suggested to use the averages $(M_i^C, M_i^T) = (\bar{Y_i}^C, \bar{Y_i}^T)$ in case of normally distributed outcomes, but proposed a paired t-test instead for a meta-analysis approach. As the standard errors for $\bar{Y_i}^C$ and $\bar{Y_i}^T$ are not the same, applying the paired t-test on $(\bar{Y_i}^C, \bar{Y_i}^T)$ contrasts with

a weighted average approach on the effect size $\bar{Y_i}^C - \bar{Y_i}^T$. Even though the pooled effect size through the paired t-test proposed by Hussey and Hughes is still correct, it is not as efficient as the weighted average on the effect sizes.[6]

WITH PERIOD EFFECTS

In case of the presence of period effects, outcomes cannot be summarized into a control-treatment pair for each clusters. This means that the analysis should be conducted on M_{ij} instead of M_i where M_{ij} represents a summary of $Y_{ij1}, \cdots Y_{ijk_{ij}}$. There are two approaches. In the first approach, variations between clusters are treated as nuisances and a marginal model with generalized estimating equations [31] can be applied. The cluster functions as the unit with repeated observations and both period and treatment enter the analysis as fixed effects. The focus of the approach is the inferences of the fixed effects averaged over the clusters.[24] The second approach is to consider a subject-specific mixed effects model with clusters as random effect and period and treatment again as fixed effects.[10, 20]

The relative merits of the two methods are well-discussed in literature.[22, 28, 30] In general, the marginal model approach does not rely on the assumption of the correlation structure and is robust against misspecification. However, when the number of clusters is small, the empirical "sandwich" estimator [19, 22, 29] used in the model underestimates the true (co)variances of the parameters and the Wald-type test is subject to inflated Type-I error. On the other hand, the mixed effects model approach is sensitive to the specification of the covariance structure and the treatment effect is more difficult to interpret on a population level.

Incorporating period effects in both approaches will be further discussed in the individual-level analysis section, since they are rather similar.

3.2.2. INDIVIDUAL-LEVEL ANALYSIS

The majority of the stepped wedge trials use an individual-level approach as their primary analysis methods. When there is no period effect, standard statistical models can be applied taking into account the possible correlations within clusters. However, in case of period effects, the analysis of data collected within a stepped wedge trial will be more complicated due to the time-dependent nature of the treatment switches. Though model specification is usually trial-specific, several general points can be made. Considering a generalized linear mixed model, there are two approaches to take period into account. Either, period can be included in the model as a categorical variable and then a piecewise constant effect for each period can be assumed. Alternatively, a functional form can be specified for the period effects by considering period as a continuous variable and a "growth" model can be specified. Both approaches has its own benefits and drawbacks. The piecewise constant approach is supposed to be less precise, but more flexible and less sensitive to misspecifications of the period effects. Whilst the growth model will be more precise, it is deemed to be problematic when an incorrect functional form is chosen for the period effects.

Another challenge in the analysis of data collected within a stepped wedge trial is the interaction between treatment with periods. This is infrequently discussed in current literature.[13] One of the main issues with interaction of treatment and period is that there is no intervention at the first period and no control at the last period in a stepped wedge trial. At least in the last period, a period-specific treatment effect is always accompanied by a period effect and therefore it is not identifiable. Treatment-period interaction can only be assessed for the periods that contain both intervention and control treatment. This has direct consequences for the parameter estimation. If there exists treatment-period interaction which is not taking into account, the estimated treatment and

period effects will be biased (as we will see later).

In case the treatment-period interactions are included in the model, then a fully parametrized model will not be identifiable. Consider a full model parametrized as in the Table 3.1 for a stepped wedge design with 5 periods and 4 switch moments. The mean response at each cell is expressed in terms of combinations of a general mean μ, period effect b_j, overall treatment effect θ, and a period-specific treatment effect δ_j as a difference with respect to the overall treatment effect. There are 8 unique cells but 11 parameters are specified.

Table 3.1 | Visualization of the full parametrization with treatment-period interaction for a stepped wedge design with 5 periods and 4 switch moments

Switch	Period 1	Period 2	Period 3	Period 4	Period 5
Switch 1	$\mu + b_1$	$\mu + b_2 + \theta + \delta_2$	$\mu + b_3 + \theta + \delta_3$	$\mu + b_4 + \theta + \delta_4$	$\mu + b_5 + \theta + \delta_5$
Switch 2	$\mu + b_1$	$\mu + b_2$	$\mu + b_3 + \theta + \delta_3$	$\mu + b_4 + \theta + \delta_4$	$\mu + b_5 + \theta + \delta_5$
Switch 3	$\mu + b_1$	$\mu + b_2$	$\mu + b_3$	$\mu + b_4 + \theta + \delta_4$	$\mu + b_5 + \theta + \delta_5$
Switch 4	$\mu + b_1$	$\mu + b_2$	$\mu + b_3$	$\mu + b_4$	$\mu + b_5 + \theta + \delta_5$

To solve the identifiability problem, it is required to eliminate 3 parameters by setting these parameters to zero in the model or put constrains on them. One possible specification is described as follow. First, for the cells under the control, there are 4 unique cells and 5 specified parameters (one mean μ and 4 period effect b_1, b_2, b_3 and b_4). Since b_1 and μ can not be estimated separately, we choose to set $b_1 = 0$. Given that, μ can be estimated from the cells of period 1. Once μ is estimable, one can also estimate b_2, b_3, and b_4 for the cells without treatment at period 2, 3 and 4. Considering the cells under the treatment, there are now 4 unique cells with 6 parameters. As in the last period, period 5, the treatment effect δ_5 is always accompanied by the period effect b_5, one could elect to set b_5 to zero. Then, there are still 3 unique cells left in period 2, period 3 and period 4 with four unknown parameters. Thus it is necessary to eliminate one more. A logical choice would be to set δ_2 to zero. Since it is the first

period to observe a treatment effect, it might be reasonable to consider this period as a reference level. The above mentioned choices would then yield a system of identifiable parameters as is shown in Table 3.2.

Table 3.2 | Visualization of the identifiable parametrization with treatment-period interaction for a stepped wedge design with 5 periods and 4 switch moments

Switch	Period 1	Period 2	Period 3	Period 4	Period 5
Switch 1	μ	$\mu+b_2+\theta$	$\mu+b_3+\theta+\delta_3$	$\mu+b_4+\theta+\delta_4$	$\mu+\theta+\delta_5$
Switch 2	μ	$\mu+b_2$	$\mu+b_3+\theta+\delta_3$	$\mu+b_4+\theta+\delta_4$	$\mu+\theta+\delta_5$
Switch 3	μ	$\mu+b_2$	$\mu+b_3$	$\mu+b_4+\theta+\delta_4$	$\mu+\theta+\delta_5$
Switch 4	μ	$\mu+b_2$	$\mu+b_3$	$\mu+b_4$	$\mu+\theta+\delta_5$

However, setting b_5 to zero is essentially assuming that period 1 and period 5 have the same effect while other periods in between having different effects. This is a rather obscured assumption to make in practice. As an alternative, we could elect to put constrains on the function form of the treatment-period interaction effects. For instance, it is sometimes reasonable to assume that the differences between different period-specific treatment effects are on average zero, namely heterogeneity of the treatment effect. Such assumption can be reflected by restricting our model using $\delta_2+\delta_3+\delta_4+\delta_5=0$. This is the same thing as setting $\delta_5=-(\delta_2+\delta_3+\delta_4)$. Note that since we have already set δ_2 zero, this is equivalent to $\delta_5=-(\delta_3+\delta_4)$. In this case, it might be preferable to consider treatment as random across period instead of using the proposed parametrization for treatment-period interaction terms. On the other hand, in certain trials, it is expected that treatment effect would improve/deteriorate over periods with a linear trend, it is then possible to assume $\delta_3-\delta_2=\delta_4-\delta_3=\delta_5-\delta_4=\Delta$, namely a constant increment for the treatment-period interactions.

3.3. SIMULATION

3.3.1. SIMULATION AND ANALYSIS

To demonstrate the points discussed in the method section, a simulation study was conducted. First of all, a cross-sectional stepped wedge design with 20 clusters, 4 switch points (5 clusters per switch points), and 5 time periods were considered. A cross-sectional design means that participants will not be followed during the trial and at each period new participants will enter the trial. For each cluster, 100 patients were simulated at each period with a binary outcome using the following generalized linear mixed model:

$$\text{logit}(\pi_{ijk}) = \mu + a_i + b_j + (\theta + \delta_j) \cdot x_{ij}$$

where π_{ijk} is the probability of experiencing the event for patient k at period j in cluster i, μ is the intercept at baseline (or the mean at the first period), a_i is a random effect for cluster i sampled from $N(0, \sigma_c^2)$, b_j is an effect of the jth period, $\theta + \delta_j$ is a period-specific treatment effect consisting of the overall treatment effect θ and δ_j the difference with respect to the overall treatment effect, and x_{ij} is the treatment indicator ($x_{ij} = 1$ means under the intervention and 0 otherwise). The variance σ_c^2 of the random effect a_i was set to 0.25.

We first considered two scenarios without treatment period interaction effects ($\theta = -0.2$ and $\delta_j = 0$ for all periods). In scenario I, we assumed that all b_j's are equal to 0 (no period effects) and for scenario II we assumed a linear trend in the period effects ($b_j = 0.2j$). Furthermore, we also considered two scenarios where the treatment-period interaction is both not zero. In these scenarios, we incorporated the same period effect as scenario II with $b_j = 0.2j$. In scenario III, a linear treatment-period interaction effect is considered with $\theta = -0.2$, $\delta_j = -0.3j$, and in

scenario IV the treatment-period interaction is considered as a random effect with $\theta = -0.2$ and δ_j sampled from a normal distribution $N(0, 0.25)$. A summary of the four simulation settings is provided in Table 3.3

Table 3.3 | Summary of the four simulation scenarios

Scenario	Period effect	Treatment period interaction	Average treatment effect
I	0	0	-0.2
II	$b_j = 0.2j$	0	-0.2
III	$b_j = 0.2j$	$\theta_j = -0.2 - 0.3j$	-1.25
IV	$b_j = 0.2j$	$\theta_j \sim N(-0.2, 0.25)$	-0.2

For all scenarios, the data was analyzed at both cluster and individual level. For the cluster-level approach, a meta-analysis approach was first considered by treating each cluster as a sub-study and we applied the Mantel-Haenszel method for the overall odds ratio. Secondly, we used a marginal model with generalized estimating equation on the aggregated event counts at cluster-period level using the binomial distribution and treat period as a categorical variable. Furthermore, a generalized linear mixed model was fitted to the aggregated data. At individual level, three different generalized linear mixed models were fitted to the simulated data. First, we used the variance component model from Hussey and Hughes which does not include the treatment-period interaction term. Secondly, we fitted the Hussey and Hughes model with additional terms for the treatment-period interactions. In addition, we fitted the Hussey and Hughes model with a constant increment in period-specific treatment effects Δ instead of the interaction term. Furthermore, a generalized linear mixed model which considers the treatment-period interaction term as a random effect is also included. Finally, we used a linear growth model by treating the period as continuous and with a slope dependent on treatment. For all models except for the Mantel-Haenszel method, clusters were considered as random as well. Additional hypothesis testing of the treatment-period interactions was made for models that take into

account the interactions based on Type III test.

Mean and standard deviation of the parameter estimations and their empirical coverage probabilities were summarized from 2000 simulations. ALL simulations were conducted in SAS® 9.4. Mantel-Haenszel estimator of the odds ratio was computed via PROC FREQ. The Greenland and Robins variance estimator for $ln(OR_{MH})$ was used to compute the confidence intervals of the Mantel-Haenszel estimates of the common odds ratios. For the marginal model with generalized estimating equations, PROC GENMOD was used and coverage probabilities were based on the Wald-type confidence intervals. For the generalized linear mixed models, PROC GLIMMIX was used with the coverage probabilities derived based on the t-type confidence intervals and the denominator degrees of freedom calculated by the default containment method. Wherever applicable, period 1 is always considered as the reference category in the analysis. For other variables, we followed the default parametrizations of the software package.

3.4. RESULTS

Simulation results of each method under the four scenarios are shown in Table 3.4 - Table 3.11, respectively. Due to the setting of the simulated stepped wedge design, the information with regarding to the treatment-period interactions can only be drawn from the three periods that have both treatment and control. Thus the inferences of the variance components of the interaction terms in the generalized linear mixed model with random treatment-period interactions were highly unreliable and is therefore omitted in the table. In scenario I, namely when there is no period effect, the cluster-level approaches all produced unbiased estimates of the treatment effect. It should be noted that since the marginal model had a population-average treatment effect interpretation, the cor-

responding mean estimation for the subject-specific treatment effect is approximately $-0.1884\sqrt{1+0.346\sigma_c^2} = -0.1964$.[18] This correction also explains the lower coverage probabilities. Individual level approaches all had unbiased estimations of the treatment effect and nominal coverage probabilities. When a secular trend was introduced into the data generation process (scenario II), the Mantel-Haenszel approach, that did not take into account the period effects, produced biased estimates of the treatment effect. Its results are therefore not presented in all scenarios with period effects. All other models had unbiased estimations of the treatment effect and the period effects. Except for the random interaction model, all other models had nominal coverage probabilities. The random interaction model, on the other hand, had too conservative coverage probabilities.

Table 3.4 | Mean, standard deviation and the empirical coverage probabilities of different methods for scenario (I): No period effect and no treatment-period interaction (Originally estimated intercept, period effects and treatment-period interaction terms from various models were suppressed for compactness)

Model	Parameter	True value	Mean estimates	Standard deviation	Empirical coverage
Mantel-Haenszel	Odds ratio	0.8187	0.8214	0.0406	95.00%
Cluster: Marginal	Treatment	-0.2	-0.1884	0.0781	91.50%
Cluster: Mixed	Treatment	-0.2	-0.1974	0.0818	94.74%
H&H (no interaction)	Treatment	-0.2	-0.1972	0.0817	94.35%
H&H (interaction)	Treatment	-0.2	-0.1986	0.1275	96.05%
Constant increment	Treatment	-0.2	-0.1963	0.1177	95.45%
Growth model	Treatment	-0.2	-0.2017	0.1561	95.00%

In scenario III when a linear treatment-period interaction effect was introduced, all models that do not take into account the interactions estimated the parameters with biases. The true value of the treatment effect was taken as the average of the four period-specific treatment effects among period 2 to 5. However, it was peculiar that the models without the interaction term were not able to estimate the average treatment effect as it would for the parallel group design situation. Apparently the average

Table 3.5 | Bias, standard deviation (SD) and the empirical coverage probabilities (CP) of three different methods without interactions for scenario (II): Period effect and no treatment-period interaction.

	Cluster: Marginal		Cluster: Mixed		H&H Model	
Parameters*	bias (SD)	CP	bias (SD)	CP	bias (SD)	CP
Treatment	0.0073 (0.0915)	92.10%	0.0001 (0.0942)	95.49%	0.0000 (0.0941)	94.95%
Intercept	0.0620 (0.1179)	90.20%	0.0066 (0.1216)	96.14%	0.0064 (0.1216)	96.15%
Period 2	0.0097 (0.0787)	92.35%	0.0025 (0.0816)	94.84%	0.0025 (0.0817)	94.40%
Period 3	0.0155 (0.0915)	90.00%	0.0010 (0.0947)	94.69%	0.0010 (0.0947)	94.25%
Period 4	0.0237 (0.1071)	89.60%	0.0024 (0.1104)	95.04%	0.0024 (0.1104)	94.55%
Period 5	0.0271 (0.1231)	90.95%	0.0007 (0.1267)	95.34%	0.0007 (0.1266)	94.90%

* Originally estimated treatment-period interaction terms were suppressed for compactness

Table 3.6 | Bias, standard deviation (SD) and the empirical coverage probabilities (CP) of three different methods with interactions for scenario (II): Period effect and no treatment-period interaction.

	H&H model with interaction		Constant increment		Growth Model	
Parameters*	bias (SD)	CP	bias (SD)	CP	bias (SD)	CP
Treatment	0.0044 (0.1590)	94.80%	0.0015 (0.1309)	95.30%	0.0048 (0.1768)	95.45%
Intercept	0.0064 (0.1216)	96.10%	0.0064 (0.1216)	96.10%	0.0064 (0.1360)	96.15%
Period 2	0.0022 (0.0869)	94.80%	0.0025 (0.0855)	94.60%	0.0008 (0.0405)	94.50%
Period 3	0.0008 (0.1051)	94.70%	0.0008 (0.0947)	94.25%		
Period 4	0.0019 (0.1487)	95.55%	0.0000 (0.1373)	95.10%		
Period 5	0.0050 (0.1800)	94.75%	0.0050 (0.2400)	95.10%		

* Originally estimated treatment-period interaction terms were suppressed for compactness.

treatment effect can no longer be estimated by these models without interaction term for stepped wedge designs. It is worth mentioning that the default parametrization of the software package for the Hussey and Hughes model with treatment-period interaction was different from the one described in Table 3.2. It had taken $\delta_4 = 0$, and $\delta_5 = 0$ which means that the estimated treatment effect now has a interpretation of $\theta + \delta_4$ which has true value of -1.4. Furthermore, since δ_5 is set to 0 as well, this is equivalent to assume that $\theta + \delta_5 = \theta + \delta_4$. Consequently, the estimated effect of period 5 had a bias of $\delta_4 - \delta_5$ compared to the true value of b_5. On the other hand, the other two interaction terms shown in the results were unbiased estimations of the treatment effect differences between period 2 (resp. period 3) and period 4: $\delta_2 - \delta_4$ (resp. $\delta_3 - \delta_4$) with its true value being 0.6 (resp. 0.3). In addition, the constant increment model also produced unbiased estimations of the parameters including the linear increment of the treatment effects. Finally, growth model had unbiased estimates with nominal coverage probabilities as well.

Table 3.7 | Bias, standard deviation (SD) and the empirical coverage probabilities (CP) of three different methods without interactions for scenario (III): Period effect and linearly increasing treatment-period interaction.

	Cluster: Marginal		Cluster: Mixed		H&H Model	
Parameters	bias (SD)	CP	bias (SD)	CP	bias (SD)	CP
Treatment	0.2299 (0.0802)	22.40%	0.1808 (0.0832)	41.84%	0.1809 (0.0832)	40.60%
Intercept	0.0640 (0.1182)	89.85%	0.0052 (0.1215)	96.25%	0.0053 (0.1216)	96.25%
Period 2	0.0734 (0.0798)	83.00%	0.0836 (0.0822)	81.98%	0.0835 (0.0823)	81.15%
Period 3	0.0350 (0.0875)	90.85%	0.0181 (0.0894)	94.89%	0.0183 (0.0895)	94.40%
Period 4	0.2876 (0.0970)	21.40%	0.2720 (0.1000)	21.97%	0.2720 (0.1000)	21.10%
Period 5	0.6371 (0.1075)	00.00%	0.6296 (0.1126)	00.00%	0.6297 (0.1126)	00.00%

In scenario IV, cluster-level marginal model, cluster-level mixed effects model, Hussey and Hughes model, and the constant increment model all had unbiased estimations for period 2, 3, and 4 but their estimations of period 5 and average treatment effect were biased and the coverage probabilities of all the parameters, except for the intercept, were too lib-

Table 3.8 | Bias, standard deviation (SD) and the empirical coverage probabilities (CP) of three different methods with interactions for scenario (III): Period effect and linearly increasing treatment-period interaction.

	H&H model with interaction		Constant increment		Growth Model	
Parameters	bias (SD)	CP	bias (SD)	c.p.	bias (SD)	CP
Treatment	0.0004 (0.1505)	95.50%	0.0055 (0.1198)	95.30%	0.0028 (0.1555)	94.75%
Intercept	0.0055 (0.1216)	96.15%	0.0055 (0.1217)	96.15%	0.0056 (0.1357)	96.40%
Period 2	0.0024 (0.0862)	95.30%	0.0025 (0.0844)	95.30%	0.0007 (0.0401)	94.40%
Period 3	0.0009 (0.1035)	94.75%	0.0015 (0.0911)	94.95%		
Period 4	0.0022 (0.1486)	94.95%	0.0008 (0.1356)	95.00%		
Period 5	0.2983 (0.1703)	55.30%	0.0025 (0.2267)	95.25%		
Period2*trt	0.0060 (0.1867)	95.70%	Δ: 0.0022 (0.0927)	95.65%	0.0004 (0.0481)	95.10%
Period3*trt	0.0031 (0.1776)	95.50%				
Period4*trt	0 (N.A.)	N.A.				
Period5*trt	0 (N.A.)	N.A.				

eral. This is probably caused by the inflated variations of the estimations. The Hussey and Hughes model with interaction terms produced unbiased estimations of all parameters with close to nominal coverage probabilities for the effects of period 2, 3, and 4. However, the coverage probabilities for period 5, the treatment effect, and the interaction terms were still anti-conservative. Furthermore, the growth model estimated the parameters correctly. However, its coverage probabilities for the treatment effect and the treatment-period interaction were too liberal.

Table 3.9 | Bias, standard deviation (SD) and the empirical coverage probabilities (CP) of three different methods without interactions for scenario (IV): Period effect and random treatment-period interaction.

	Cluster: Marginal		Cluster: Mixed		H&H Model	
Parameters	bias (SD)	CP	bias (SD)	CP	bias (SD)	CP
Treatment	0.0365 (0.2914)	49.40%	0.0276 (0.2975)	46.39%	0.0280 (0.2977)	45.75%
Intercept	0.0514 (0.1200)	91.20%	0.0739 (0.1231)	95.68%	0.0026 (0.1231)	95.65%
Period 2	0.0005 (0.1287)	78.10%	0.0029 (0.1338)	78.11%	0.0028 (0.1342)	77.30%
Period 3	0.0104 (0.2060)	66.00%	0.0152 (0.2099)	62.40%	0.0153 (0.2102)	61.75%
Period 4	0.0007 (0.3602)	47.55%	0.0063 (0.3620)	46.59%	0.0068 (0.3621)	45.80%
Period 5	0.0378 (0.5657)	34.50%	0.0461 (0.5795)	33.79%	0.0466 (0.5792)	33.15%

For the hypothesis testing of the treatment-period interaction terms,

Table 3.10 | Bias, standard deviation (SD) and the empirical coverage probabilities (CP) of three different methods with interactions for scenario (IV): Period effect and random treatment-period interaction.

	H&H model with interaction		Constant increment		Growth Model	
Parameters	bias (SD)	CP	bias (SD)	CP	bias (SD)	CP
Treatment	0.0116 (0.5203)	45.75%	0.0046 (0.4375)	42.85%	0.0328 (0.8852)	33.05%
Intercept	0.0028 (0.1230)	95.80%	0.0028 (0.1230)	95.65%	0.0025 (0.1383)	96.20%
Period 2	0.0009 (0.0857)	94.75%	0.0002 (0.1014)	90.85%	0.0007 (0.0403)	94.65%
Period 3	0.0008 (0.1034)	95.20%	0.0080 (0.2069)	62.95%		
Period 4	0.0032 (0.1486)	95.00%	0.0173 (0.2646)	69.90%		
Period 5	0.0302 (0.7291)	37.90%	0.0693 (0.9586)	37.25%		
Period2*trt	0.0274 (0.7509)	41.55%	Δ: 0.0184 (0.3735)	41.30%	0.0025 (0.2381)	33.30%
Period3*trt	0.0337 (0.7366)	41.65%				
Period4*trt	0 (N.A.)	N.A.				
Period5*trt	0 (N.A.)	N.A.				

the percentage of the simulations that produced significant results from the three models, namely the Hussey and Hughes model, the constant increment model and the growth model, are shown in Table 3.11. When there is no interactions between treatment and period, all three models had Type I errors less than 5%. In scenario III, all three models had larger than 80% power. The growth model had the highest power of 100.00% and the Hussey and Hughes model with interactions had lowest power of 83.35%. The constant increment model performed in between. In scenario IV with random treatment-period interactions, the Hussey and Hughes model still maintained a power of 80.25% while the other two models both had significantly worsened powers.

Table 3.11 | The percentages of results with p-value smaller or equal to 0.05 from the hypothesis testing of treatment-period interaction effects for the three models with interactions.

Scenario	H&H model	Constant increment	Growth Model
I	04.30%	03.70%	04.60%
II	04.35%	04.50%	04.85%
III	83.35%	90.00%	100.00%
IV	80.25%	58.70%	66.70%

3.5. DISCUSSION

In the present paper, we discussed some practical issues in terms of analyzing data for a stepped wedge design at cluster level and individual level. In general, without stringent assumptions on the absence of period effects and period-treatment interactions, standard statistical methods are frequently insufficient and leads to possibly incorrect interpretations and conclusions.

Indeed in classic parallel setting, one would still expect the frequently used models without period-treatment interaction such as the Hussey and Hughes model to be able to estimate the average treatment effect. However, this is no longer the case under the stepped wedge setting. This should raise a lot of attention about the consequences of fitting a model without the interaction but still interpret the results as in the parallel design. Therefore, it is crucial to assess the treatment-period interaction terms first. According to the simulation results, we recommend to use the generalized linear mixed model of Hussey and Hughes with the inclusion of treatment-period interaction to investigate the differences between the specific treatment effects at different periods since it has consistent power to detect the interactions. Even though the treatment effect at the last period is not estimated unbiasedly in this model, the estimations of other interaction terms can be used to aid the judgement of whether there exists a treatment-period interaction. Furthermore, it provides opportunities to explore the specific form of the treatment-period interactions which allows a correct parametrization/model to be used. For instance, if the interaction term is linear, the constant increment model proposed in the paper or a growth model would be preferred. On the other hand, if the interactions are truly random, there is at the moment no models that can consistently estimate the parameters with nominal coverage probabilities. The random interaction model might be a good candidate

if there is sufficient numbers of periods that can provide information of the interactions. Nevertheless, it is still unappealing to assume that the interactions of treatment and period is random but both treatment and period are not. A better model to deal with scenario IV for stepped wedge design is of great interests for investigations.

It is noteworthy that the parametrization of the commonly used statistical software such as the case in SAS is not the same as the ones proposed in the presented paper. On the other hand, it is straightforward to include interactions in a growth model but the problem becomes more complex when the three-way interactions between treatment, period and cluster is considered. Due to the limited space, this problem was not studies in the present paper and further investigations is needed.

Meta-analysis methods are very strong and are serious analysis candidates when period effects are non-existent. Further benefits of applying meta-analysis methods is the ability to quantifying and testing heterogeneity of the effect sizes [5, 14] which is not often considered in stepped wedge designs. Rejection of the test implies the presence of heterogeneity of the population effects. By using random effects instead of fixed effect meta-analysis methods one can account for this in the analysis.[26, 27]

Overall, period effect, correlation within clusters and treatment heterogeneities are three important questions to consider prior to the analysis of the data in the stepped wedge design.

REFERENCES

[1] Barker D, McElduff P, D?ste C, Campbell M (2016) Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC Med Res Methodol 16(1):1

[2] Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, et al (2015) Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials 16(1):1

[3] Brown CA, Lilford RJ (2006) The stepped wedge trial design: a system-

atic review. BMC Med Res Methodol 6(1):1

[4] Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M (2000) Analysis of cluster randomized trials in primary care: a practical approach. Family Practice 17(2):192, DOI 10.1093/fampra/17.2.192

[5] Cochran WG (1954) The combination of estimates from different experiments. Biometrics 10(1):101–129

[6] Cooper H, Hedges LV, Valentine JC (2009) The handbook of research synthesis and meta-analysis. Russell Sage Foundation

[7] Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR (2015) Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials 16(1):352

[8] Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ, Fielding KL (2015) Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. Trials 16(1):358

[9] DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7(3):177–188

[10] Girling AJ, Hemming K (2016) Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Stat Med 35(13):2149–2166, DOI 10.1002/sim.6850

[11] Hemming K, Taljaard M (2016) Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. J Clin Epidemiol 69:137–146

[12] Hemming K, Haines T, Chilton P, Girling A, Lilford R (2015) The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 350:h391

[13] Hemming K, Taljaard M, Forbes A (2017) Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials 18(1):101

[14] Higgins J, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Stat Med 21(11):1539–1558

[15] de Hoop E, van der Tweel I, van der Graaf R, Moons KG, van Delden JJ, Reitsma JB, Koffijberg H (2015) The need to balance merits and limitations from different disciplines when considering the stepped wedge cluster randomized trial design. BMC Med Res Methodol 15(1):1

[16] van Houwelingen HC, Zwinderman KH, Stijnen T (1993) A bivariate approach to meta-analysis. Stat Med 12(24):2273–2284

[17] van Houwelingen HC, Arends LR, Stijnen T (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 21(4):589–624

[18] Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA (1998) Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. Am J Epidemiol 147(7):694–703

[19] Huber PJ (1967) The behavior of maximum likelihood estimates under non-

standard conditions. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Berkeley, CA, vol 1, pp 221–233

[20] Hussey MA, Hughes JP (2007) Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 28(2):182–191

[21] Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W (2012) Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. J Clin Epidemiol 65(12):1249–1252

[22] Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika pp 13–22

[23] Mdege ND, Man MS, Taylor CA, Torgerson DJ (2012) There are some circumstances where the stepped-wedge cluster randomized trial is preferable to the alternative: no randomized trial at all. response to the commentary by Kotz and colleagues. J Clin Epidemiol 65(12):1253

[24] Scott JM, Juraska M, Fay MP, Gilbert PB, et al (2014) Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Stat Methods Med Res p 0962280214552092

[25] Thompson JA, Fielding KL, Davey C, Aiken AM, Hargreaves JR, Hayes RJ (2017) Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. Stat Med

[26] Thompson SG (1994) Why sources of heterogeneity in meta-analysis should be investigated. BMJ 309(6965):1351

[27] Thompson SG, Sharp SJ (1999) Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 18(20):2693–2708

[28] Verbeke G (2005) Models for Discrete Longitudinal Data. Springer Series in Statistics. Springer

[29] White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: Journal of the Econometric Society pp 817–838

[30] Zeger SL, Liang KY (1992) An overview of methods for the analysis of longitudinal data. Stat Med 11(14-15):1825–1839

[31] Zeger SL, Liang KY, Albert PS (1988) Models for longitudinal data: a generalized estimating equation approach. Biometrics pp 1049–1060

[32] Zhan Z, van den Heuvel ER, Doornbos PM, Burger H, Verberne CJ, Wiggers T, de Bock GH (2014) Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. J Clin Epidemiol 67(4):454–461

4

INTENSIFIED FOLLOW-UP IN COLORECTAL CANCER PATIENTS USING FREQUENT CARCINO-EMBRYONIC ANTIGEN (CEA) MEASUREMENTS AND CEA-TRIGGERED IMAGING: RESULTS OF THE RANDOMIZED "CEAWATCH" TRIAL

C. J. Verberne, Z. Zhan
E. R. van den Heuvel, I. Grossmann
P. M. Doornbos, K. Havenga
J. Klaase, H. C. J. van der Mijle
B. Lamme, K. Bosscha
P. Baas, B. van Ooijen
G. Nieuwenhuijzen, A. Marinelli
E. van der Zaag, D. Wasowicz
G. H. de Bock, T. Wiggers

This chapter has been published in European Journal of Surgical Oncology, Volume **41**, Issue 9, (2015) [34].

ABSTRACT

Aim: The value of frequent Carcino-Embryonic Antigen (CEA) measurements and CEA-triggered imaging for detecting recurrent disease in colorectal cancer (CRC) patients was investigated in search for an evidence-based follow-up protocol.

Methods: This is a randomized-controlled multicenter prospective study using a stepped-wedge cluster design. From October 2010 to October 2012, surgically treated non-metastasized CRC patients in follow-up were followed in eleven hospitals. Clusters of hospitals sequentially changed their usual follow-up care into an intensified follow-up schedule consisting of CEA measurements every two months, with imaging in case of two CEA rises. The primary outcome measures were the proportion of recurrences that could be treated with curative intent, recurrences with definitive curative treatment outcome, and the time to detection of recurrent disease.

Results: 3223 patients were included; 243 recurrences were detected (7.5%). A higher proportion of recurrences was detected in the intervention protocol compared to the control protocol (OR = 1.80; 95%-CI: 1.33–2.50; p = 0.0004). The proportion of recurrences that could be treated with curative intent was higher in the intervention protocol (OR = 2.84; 95%-CI: 1.38–5.86; p = 0.0048) and the proportion of recurrences with definitive curative treatment outcome was also higher (OR = 3.12, 95%-CI: 1.25–6.02, p-value: 0.0145). The time to detection of recurrent disease was significantly shorter in the intensified follow-up protocol (HR = 1.45; 95%-CI: 1.08–1.95; p = 0.013).

Conclusion: The CEAwatch protocol detects recurrent disease after colorectal cancer earlier, in a phase that a significantly higher proportion of recurrences can be treated with curative intent.

4.1. INTRODUCTION

After curative surgical resection of colorectal cancer (CRC) and termination of adjuvant treatments, patients are offered a follow-up program consisting of imaging, laboratory measurements, and physical examination to detect recurrent disease as early as possible. The use of an intensive follow-up regime results in a modest but statistically relevant improvement in survival compared with a minimal strategy [4, 7, 27, 29, 31], but this conclusion is based on older studies (inclusion period 1983–2001) and most studies were considered to be of poor quality.[25] The survival gain of intensive protocols is considered to be the effect of detecting recurrences at an earlier stage, associated with a higher rate of curative treatment.[19, 25]

Routine imaging with ultrasound of the liver is advised twice yearly for the first three years and once annually in years 4 and 5 in the Dutch national guideline (2008) (`www.oncoline.nl`). Computed Tomography (CT) scanning is an alternative with higher sensitivity [21], but it is costly and has the potential disadvantages of radiation damage [3] and false positive findings.[11]

The tumour marker Carcino-Embryonic Antigen (CEA) has long been known to be important in signalling recurrent disease in CRC.[30] Intensive follow-up schedules including CEA measurements are correlated with better survival than schedules not using CEA measurements [19], and serial measurements of CEA are recommended in colorectal cancer follow-up in all international guidelines.[6, 23, 33] The rise and doubling time of CEA rather than the absolute value are sensitive in signalling recurrent disease.[30, 38] CEA is cheap and available, but is irregularly used in follow-up and has poor protocol adherence.[8, 10] No studies of serial CEA measurements and imaging steps in response to significant CEA rise, with special attention to reasonable sensitivity in combination with good

specificity, have been performed so far.

There is a need for an evidence-based follow-up guideline defining the optimal frequency and implications of imaging and CEA measurements. A phase-2 trial with monthly CEA measurements showed both high sensitivity and specificity for detection of recurrences using serial CEA rises rather than absolute values.[12] Therefore, a promising solution may be frequent CEA testing with imaging triggered by a significant rise in CEA. The current study compared a new intensified follow-up schedule with care as usual in a randomized multicenter trial and aimed to assess the value of frequent CEA measurements and CEA-triggered imaging in detecting recurrent disease with curative possibilities in CRC patients.

4.2. MATERIALS AND METHODS

4.2.1. TRIAL DESIGN

This was a multicenter stepped-wedge cluster randomized (SW-CRT) trial [16, 17] conducted in 11 non-academic teaching hospitals in the Netherlands. These hospitals were randomly grouped into five clusters. Detailed explanation on the motivation of using this trial design is given by Zhan et al.[39]

In an SW-CRT, all clusters cross over from control to intervention at certain time points called switches. Instead of randomizing patients to treatment arms, randomization is used to allocate clusters to predefined switches. For the CEAwatch trial, each of the clusters was randomly switched to change from the usual follow-up schedule (control) to the intensive follow-up schedule (intervention); crossover occurred in one direction only. From October 2010 clusters switched from usual follow-up to intensive follow-up every three months one by one; the length between two consecutive switches was three months (Fig. 4.1).

Randomization was performed independently by Trial Coordination

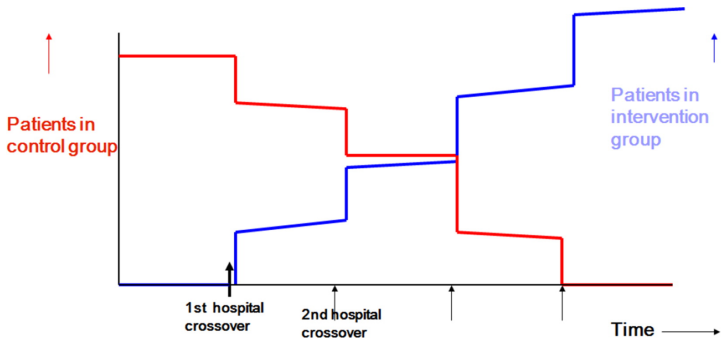

Figure 4.1 | Graphic depiction of the stepped wedge cluster randomized trial.

Center (TCC) Groningen (www.tcc.umcg.nl). CEAwatch (Netherlands Trial Register [NTR] 2182) was approved by the Medical Ethics Committee of the University Medical Centre Groningen (METc-UMCG 2010.064) and the local ethics committees of all participating centres. CEAwatch was sponsored by the Netherlands Organization for Health Research and Development and undertaken in accordance with the principles of Good Clinical Practice.

4.2.2. PARTICIPANTS

Eligible patients were patients with AJCC stage I–III CRC after R0 resection. Patients operated on from 2007 to July 2012 were included. At October 2010, patients who were already in follow-up in the participating hospitals since 2007 were included. Between October 2010 and October 2012, all new patients that entered follow-up in the participating hospitals were included and assigned to the actual protocol of that hospital.

Patients who were not medically fit for metastasectomy, patients diagnosed with other malignancies and patients with metachronous metastases at the start of the study were excluded.

4.2.3. PATIENT IDENTIFICATION AND VALIDATION

Eligible patients were identified using the diagnosis or operation code(s). At the end of patient recruitment (October 2012), eligibility of all patients was validated using the database of the Dutch Comprehensive Cancer Center (NCCC), a registry of all diagnosed malignancies based on the automated pathological archive (www.iknl.nl).

Patients' characteristics were obtained directly from the Dutch Surgical Colorectal Audit (DSCA) and stored in a password-protected database. DSCA is a national databank gathering all relevant information on surgically treated CRC patients, allowing a valid and complete registration of all CRC patients in the Netherlands (www.clinicalaudit.nl/dsca).

4.2.4. FOLLOW-UP SCHEDULES

The control or "care-as-usual" protocol consisted of the national guideline in the Netherlands in 2008 (www.oncoline.nl); an outpatient clinic visit every six months for the first three years and an annual visit in years 4 and 5. Liver ultrasound and chest X-ray were recommended at each clinic visit. CEA (half-life: 5 days) was measured every 3–6 months in the first three years and each year in the last two years.

The intervention follow-up protocol adhered to bi-monthly CEA measurements and yearly imaging in the first three years, and 3-monthly CEA measurements in the fourth and fifth years of follow-up. Outpatient clinic visits with imaging of chest and abdomen were performed annually in the first three years. In case of an increase of 20% compared with the previous CEA with CEA value >2.5 ng/ml, another blood sample was drawn four weeks later. If a consecutive rise were was observed, a CT scan of chest and abdomen was advised (Fig. 4.2). The static normal value of serum-CEA as advised by manufacturers is 2 ng/ml to 2.5 ng/ml, depending on the actual test. The coordination and monitoring of this process was supported

by an automatic computer system.[35]

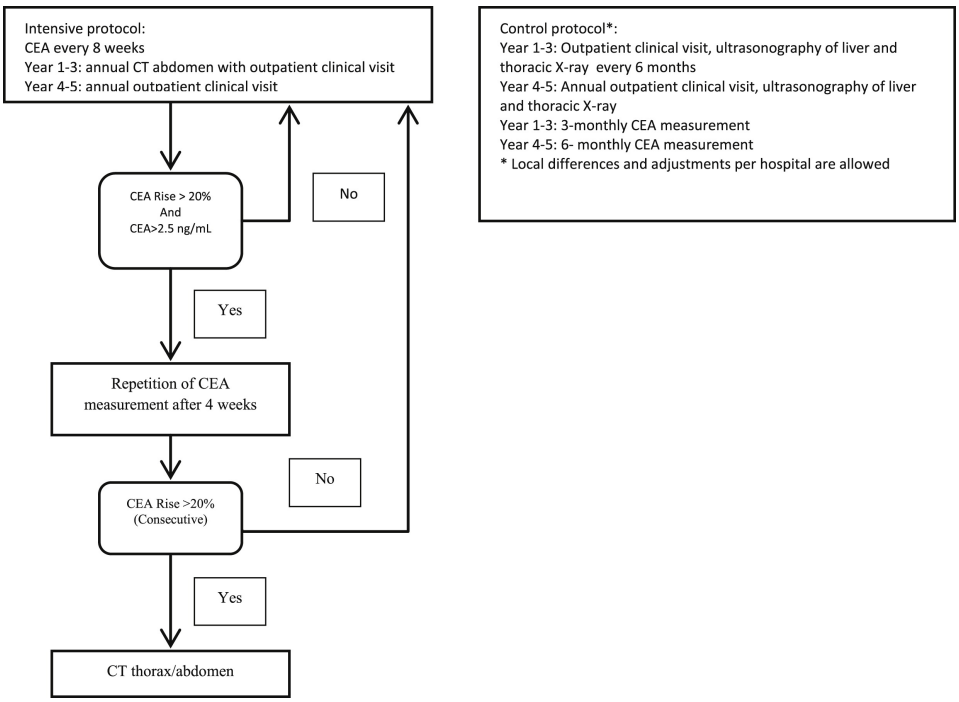

Figure 4.2 | Follow-up schedules.

4.2.5. IMPLEMENTATION

Patients entering the study before the switch were followed using the control protocol and switched to the intervention after their hospital's switch. Patients entering the study after the randomized switch of a hospital were followed using the intervention protocol only. Patients who met the inclusion criteria on October 1st, 2010, but no longer met these criteria at the switch date participated in the control protocol only. Informed consent was obtained before entering the intervention for all patients as required by the Medical Ethical Committee.

4.2.6. OUTCOMES

The primary outcome measures were the number of recurrences per follow-up arm, the proportion of recurrences that could be treated with curative intent, the proportion of recurrences with definitive curative treatment outcome (R0 resection of all recurrent disease), and the time to detection of recurrent disease.

4.2.7. POWER CALCULATION

The expected percentage of resectable recurrences was 10% in the control protocol and 25% in the intensified protocol.[1, 24] Given a significance level of 5% and a power of 80%, 115 patients with recurrent disease in both groups were needed. Given an expected recurrence rate of 25%,[22] 460 patients per group were needed. Given the cluster randomization, we assumed a correlation of 0.1 between hospitals, yielding a correction factor of 1.71.[18] Therefore, a minimum of about 800 patients per group was needed.

4.2.8. DATA ANALYSIS

Differences in patients' baseline characteristics between the care as usual follow-up and the intensified follow-up protocol were calculated using ANOVA and Chi-Square tests.

For each of the three outcomes (recurrence, recurrence with curative intent and recurrence with definite curative treatment outcome), pooled logistic regression was performed to compare the proportion of each outcome between the control follow-up protocol and the intensified follow-up due to the fact that standard statistical technique cannot address the dynamic settings of the follow-up protocol. The study duration was divided into six intervals by the five switch moments. The conditional probability of the outcome measures in each interval, given that

this did not happen prior to this interval, was modelled as the dependent variable and the follow-up protocol of each interval was modelled as the independent variable. Meanwhile generalized estimation equation (GEE) was used to allow flexible assumptions of the correlations between each interval. Odds Ratios (OR) with 95% confidence intervals (95%-CIs) were reported for the effects of the intensified follow-up protocols on the detection of recurrences, detection of recurrences treated with curative intent and recurrences with definitive treatment outcome. The Cox proportional hazard model was used to investigate the differences in time till detection of recurrent disease between the follow-up protocols. The follow-up protocols were used as a time-dependent variable since the time in follow-up was dynamic. The time from operation to participation in the study created left truncated data for a subset of patients. Stratification for hospitals was applied. The intervention effect was corrected for gender, age, AJCC stage, and location of the primary tumour and it was reported as hazard ratios (HR) with 95%-CIs. Statistical modelling was performed with SAS statistical software, version 9.3.

4.3. RESULTS

4.3.1. INCLUSIONS

From 1-1-2007 till 01-10-2010, 5604 patients from 11 hospitals with stage AJCC I–III colorectal cancer were registered by the Netherlands Cancer Registration; 118 patients were not identified in the hospitals. Of these patients, 2318 met the inclusion criteria; their follow-up data were prospectively collected from 01-10-2010. During the control period, there were 589 eligible new patients identified before the switch dates; 116 patients reached an endpoint (not fit for metastasectomy, recurrent disease before switch date, or other) during the control period. A total of 2791 patients were asked for informed consent prior to the switch dates. Of these, 1725

patients provided written informed consent. For the remaining 1066 patients, prospective data collection of follow-up data ended on the switch dates. During the intervention period, an additional 316 patients gave written informed consent to participate in the intensive follow-up protocol.

A total of 3223 patients were included. 1725 patients participated both in the control protocol and in the intervention protocol, 1182 patients participated only in the control protocol, and 316 patients participated only in the intervention protocol (Fig. 4.3).

In total, the control period comprised 2907 patients and the intervention period comprised 2041 patients. Patient's characteristics are given in Table 4.1. The differences between eligible patients who decided to participate and eligible patients who decided not to participate in the intervention protocol are shown in Table 4.2.

4.3.2. RECURRENCES

A total of 243 (7.5%) recurrences were detected during the study (Table 4.3). 104 (43%) recurrences were found while the patient participated in the control protocol and 139 (57%) recurrences were detected while the patient participated in the intervention protocol. 90 (37.0%) of all recurrences could be treated with curative intent.

The proportion of detected recurrences eligible for curative treatment during the intervention protocol was higher than in the control protocol (42.0% versus 30.0%). Further analysis with results of real pathology (treatment outcome instead of treatment intent) showed that and 70 (78%) of all detected recurrences treated with curative intent had definite curative treatment outcome based on pathology: the proportion of curative treatment outcome was also higher in the intervention than in the control (35% versus 22%).

The location of detected recurrences ($p = 0.134$), AJCC stage of the

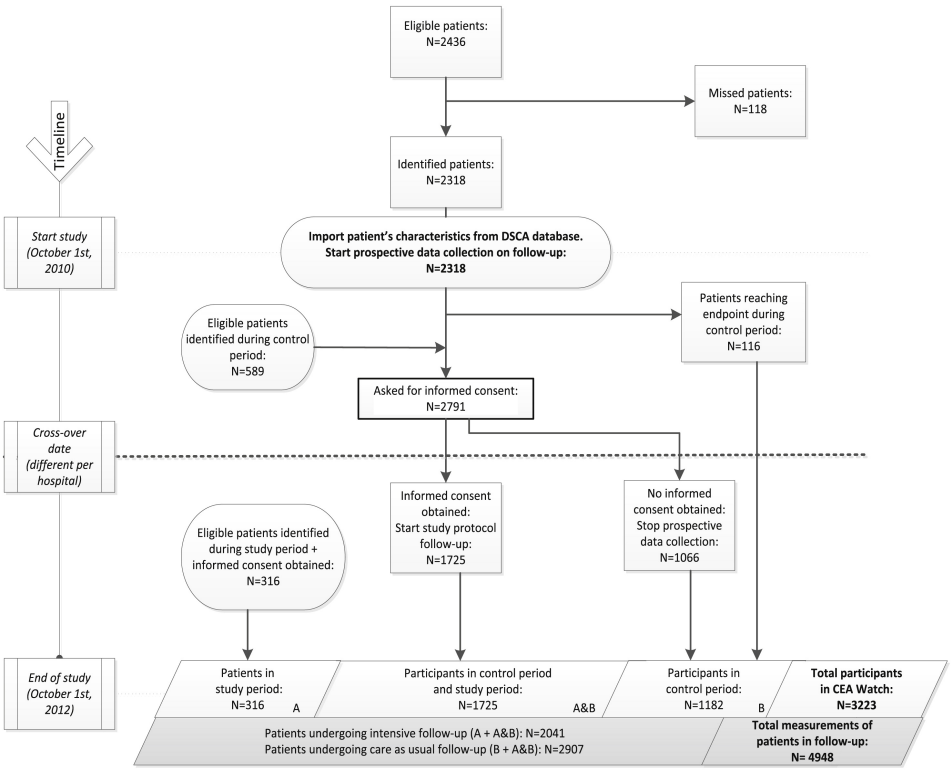

Figure 4.3 | Inclusions.

Table 4.1 | Patient's and tumour characteristics.

Characteristic	Patients only in control period	Patients in control and intervention period	Patients only in intervention period	Total	p-value[a]
Total	1182 (37%)	1725 (53%)	316 (10%)	3223	
Gender					<0.01
Male	603 (51%)	1024 (59%)	180 (57%)	1807 (56%)	
Female	579 (49%)	701 (41%)	136 (43%)	1416 (44%)	
Age at diagnosis (years)					<0.01
Median (range)	73 (26-95)	69 (30-93)	67 (29-92)	70 (26-95)	
AJCC stage[b]					<0.01
I	281 (24%)	504 (29%)	92 (29%)	887 (28%)	
II	462 (39%)	670 (39%)	137 (43%)	1269 (39%)	
III	439 (37%)	551 (32%)	87 (26%)	1077 (33%)	
Location primary tumor					0.4
Colon	754 (64%)	1068 (62%)	206 (65%)	2028 (63%)	
Rectum	428 (36%)	657 (38%)	110 (35%)	1195 (37%)	
Adjuvant chemotherapy[c]	299 (43%)	337 (49%)	55 (8%)	691 (100%)	0.04
Yes	187 (63%)	249 (74%)	45 (82%)	481 (70%)	
No	112 (37%)	88 (26%)	10 (18%)	210 (30%)	
Patients with comorbidity[d]	370 (31%)	768 (56%)	225 (17%)	1363 (100%)	0.06
None	145 (39%)	369 (48%)	95 (42%)	609 (45%)	
Minor	195 (53%)	341 (44%)	114 (51%)	650 (48%)	
Major	30 (8%)	58 (8%)	16 (7%)	104 (7%)	

[a] These p-values were calculated using ANOVA and Chi-Square tests.
[b] AJCC: American Joint Committee on Cancer.
[c] For adjuvant chemotherapy, only patients with stage III colon cancers are shown.
[d] For comorbidity, only patients with known comorbidity are shown. P-value is calculated for the group with no comorbidity versus minor or major comorbidity.

Table 4.2 | Comparison between patients deciding to participate in the intervention follow-up protocol and patients deciding not to participate in the intervention protocol.

Characteristic	All patients eligible for intervention protocol	Patients not crossing over to intervention protocol[a]	Patients crossing over to intervention protocol	p-value[b]
Total	2791	1066 (38.2%)	1725 (61.8%)	
Gender				<0.01
Male	1562 (56%)	538 (50%)	1025 (59%)	
Female	1229 (44%)	528 (50%)	701 (41%)	
Age at diagnosis (years)				<0.01
Median (range)	70 (26-95)	73 (26-95)	69 (30-93)	
AJCC stage[c]				0.44
I	767 (27%)	263 (25%)	504 (29%)	
II	1093 (39%)	423 (40%)	670 (39%)	
III	931 (34%)	380 (35%)	551 (32%)	
Location primary tumor				0.45
Colon	1744 (63%)	676 (63%)	1068 (62%)	
Rectum	1047 (37%)	390 (37%)	657 (38%)	
Adjuvant chemotherapy				0.13
Yes	737 (26%)	295 (28%)	442 (6%)	
No	2054 (73%)	771 (72%)	1283 (74%)	
Patients with comorbidity[d]	1121 (100%)	353 (32%)	768 (68%)	0.01
None	509 (45%)	140 (40%)	369 (48%)	
Minor	528 (47%)	187 (53%)	341 (44%)	
Major	84 (8%)	26 (7%)	58 (8%)	

[a] These were all the patients eligible to cross over who did not consent to cross-over to the new follow-up regimen.
[b] These p-values were calculated using ANOVA and Chi-Square tests.
[c] AJCC: American Joint Committee on Cancer.
[d] For comorbidity, only patients with known comorbidity are shown. P-value is calculated for the group with no comorbidity versus minor or major comorbidity.

Table 4.3 | Location and treatment of recurrences in control and intervention protocol (N; %).

Variable	Total	Control period	Intervention period	p-value[a]
Recurrent disease	243 (8)	104 (43)	139 (57)	<0.001
Treatment for recurrent disease				0.03
Curative	90 (37)	31 (30)	59 (42)	
Palliative	153 (63)	74 (70)	79 (58)	
Location of recurrent disease				0.13[b]
Liver	89 (36)	41 (39)	48 (35)	
Local recurrence	44 (18)	13 (13)	31 (22)	
Lymph nodes	15 (6)	8 (8)	7 (5)	
Lung	48 (20)	17 (16)	31 (22)	
Other	24 (10)	15 (14)	9 (7)	
Combination	23 (10)	10 (10)	13 (9)	
AJCC stage primary tumour				0.98[c]
I	23 (9.5)	10 (9.6)	13 (9)	
II	89 (36.6)	36 (34.6)	53 (38)	
III	131 (53.9)	58 (55.8)	73 (53)	
Location primary tumour				0.26
Colon	145 (60)	68 (65)	77 (55)	
Rectum	98 (40)	36 (35)	62 (45)	

[a] These p-values were calculated with a logistic regression stratified for centre.

[b] This p-value was calculated by comparing recurrences in liver versus in other locations, stratified for centre.

[c] This p-value was calculated by comparing AJCC stage I and II versus III, stratified for centre.

primary tumour (p = 0.978) and the location of the primary tumour (p = 0.261) were not different in both follow-up protocols.

Pooled logistic regression showed statistically significant higher proportion of recurrences in the intervention protocol compared to the control protocol (OR = 1.80, 95%-CI: 1.33–2.50, p-value: 0.0004). The proportion of recurrences that could be treated with curative intent was also statistically significant higher in the intervention protocol (OR = 2.84, 95%-CI: 1.38–5.86, p-value: 0.0048).

The OR of recurrences with definite curative treatment outcome was also higher in the intervention protocol (OR = 3.12, 95%-CI: 1.25–6.02, p-value: 0.0145).

The time to diagnosis of recurrent disease, corrected for age, gender, AJCC stage and location of the primary tumour, and stratified by hospital using the Cox proportional hazard model, decreased with the intervention follow-up protocol as compared to the control protocol (HR: 1.45; 95%-CI: 1.08–1.95; p = 0.013). This was also shown for the recurrences treated with curative intent (HR: 1.76; 95%-CI: 1.07–2.90; p = 0.027) and recurrences with definite curative treatment outcome (HR: 6.27; 95%-CI: 3.82–10.30; p <0.0001).

4.4. DISCUSSION

In the current study including 3223 patients, it is shown that an intensified follow-up schedule with frequent CEA measurements, CEA slope analyses instead of absolute values and imaging in case of two subsequent CEA rises detects recurrences with higher rate of curable options (42% versus 30%), higher rate of definitive treatment outcome (35% versus 22%) and less time-to-detection compared to a care as usual follow-up protocol. To date there has been no randomized trial for colorectal cancer follow-up with so many participants.

Intensity of colorectal cancer follow-up schedules has been the subject of discussion for decades but in the studies performed to date both the use of CEA and imaging are heterogeneous between studies.[4, 7, 21, 27, 28, 31] All performed studies so far lack a description of a systematic plan of action in case of a CEA rise resulting in the impossibility to describe the best combination of techniques for the ideal CRC follow-up.[19] The expanding options for curing liver metastases show that intensive systematic searching for liver metastases is worthwhile.[14, 20] At least as important is the growing evidence that limited extrahepatic diseases as well as local recurrent disease are no longer an absolute contraindication for intended curative treatment.[5, 15] However, the definition of curable or resectable recurrences is difficult and differs per hospital, especially for Radio-Frequent Ablation options and stereotactic radiation therapy.[37]

An optimal follow-up schedule should detect recurrences in an early stage. The balance between false positive findings as a result of a too sensitive test reflecting normal CEA variations or not yet detectable recurrent disease and the too late detection is crucial. An analysis on older data using CEA showed a lack of survival improvement for second-look operations based on CEA rise.[32] The FACS trial, a randomized trial comparing minimal and intensive follow-up, recently confirmed that regular CEA measurements, CT scanning and CEA with CT scanning result in significantly higher rates of curable recurrences compared to minimum follow-up (resp 7.6%, 9.5%, 7.3% and 1.5).[26] However there was no survival improvement between the different follow-up protocols in this study. A recent systematic review and meta-analysis included next to the FACS trial all old studies; a modest survival improvement for intensified protocols was shown.6 However, it can be questioned whether this estimate is unbiased since the incidence of recurrences is lowering and the options for cure of recurrences are expanding, and only one recent study was included in the meta-analysis. Data from two other prospective trials

(the COLOFOL trial [36] and the GILDA trial [9], both comparing overall and disease-specific survival between different follow-up schedules) will become available.

Relatively few recurrences (7.5%) were found in the here presented study; the expected recurrence rate for AJCC stages I–III of colorectal carcinomas is about 20%.[2] In the FACS study this percentage is also lower in comparison with the older literature, namely 16%. The Dutch national guideline on routine preoperative staging with CT scan seems to result in more synchronous and less metachronous metastases.[13] Hereby the intention of the study is to cover a period of five years of follow-up and patients with a disease-free period before the schedule started were included. The prospective data collection of these patients started sometimes 2–3 year after resection, decreasing the expected recurrence rate. The total number of patients included was high enough to detect statistically significant differences.

A strong point in this study is the high data integrity, as all data on patients' and tumour characteristics were exported from a national audit which is known to be filled out for up to 97% of all colorectal cancer patients (www.clinicalaudit.nl). Data monitoring was performed through a secondary validation using the NCCC, which is the complete cancer registration in the Netherlands. Another strong point of this study was the uniformity of the intervention protocol and high adherence to the protocol. This was the result of a software-support system for the management of all patients in the intervention group, an intranet-based software system written to support clinicians working with patients in follow-up. The software support has been shown to be safe and efficient.[35]

Internationally, CT scanning is common practice and ultrasound with thoracic X-ray which seem a bit old-fashioned. However this study is performed to compare the usual follow-up with a new schedule; the study was performed during the time that the 2008 Dutch national guideline

was used and this guideline advised X-ray and ultrasound.

The SW-CRT has not previously been used for the purposes of a follow-up study. Advantages of the design are the inclusion of large patient groups in a short time period and the avoiding of in-hospital protocol contamination. On the other hand, patients participating in both follow-up protocols are always later in the intervention protocol than in the control protocol, which makes the SW-CRT not a pure RCT. Meanwhile, the incidence of recurrence tends to change over time during follow-up. Most recurrences are found in the first two years of follow-up, but retaining percentages of recurrences are seen in the years thereafter.[20] Thus, it can never be known whether the observed effects are completely due to the intervention. However, as shown in the results, the increase in resectable recurrences was not entirely due to the increases of recurrences since the effect size of the intervention is much larger for resectable recurrences.

The current study shows that an intensified protocol with CEA and assessment on CEA rise rather than absolute value detects recurrences earlier than the standard protocol, which is related to an increase in curable recurrence rate. The results advocate an intensification of CEA measurements and more frequent action at CEA rises in follow-up. The FACS study is using an absolute CEA cut-off point of 7 µg/l compared to baseline instead of slope analyses; in the discussion the authors advocate further analyses on this matter, but in the current results of this study, already addressing CEA changes, no further conclusions on this topic can be drawn. The final proof of the value and strength of this new protocol will be if the effects of the intensified CEA-based follow-up strategy will result in higher disease-specific and overall survival, with acceptable quality of life and cost-effectiveness rates.

REFERENCES

[1] Bentrem DJ, DeMatteo RP, Blumgart LH (2005) Surgical therapy for metastatic disease to the liver. Annu Rev Med 56:139–156

[2] Böhm B, Schwenk W, Hucke H, Stock W (1993) Does methodic long-term follow-up affect survival after curative resection of colorectal carcinoma? Diseases of the colon & rectum 36(3):280–286

[3] Brenner DJ, Hall EJ (2007) Computed tomography—an increasing source of radiation exposure. N Engl J Med 357(22):2277–2284

[4] Bruinvels DJ, Stiggelbout AM, Kievit J, van Houwelingen HC, Habbema JD, van de Velde CJ (1994) Follow-up of patients with colorectal cancer. A meta-analysis. Ann Surg 219(2):174

[5] Carpizo DR, D'angelica M (2009) Liver resection for metastatic colorectal cancer in the presence of extrahepatic disease. The lancet oncology 10(8):801–809

[6] Duffy M, van Dalen A, Haglund C, Hansson L, Holinski-Feder E, Klapdor R, Lamerz R, Peltomaki P, Sturgeon C, Topolcan O (2007) Tumour markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines for clinical use. Eur J Cancer 43(9):1348–1360

[7] Figueredo A, Rumble RB, Maroun J, Earle CC, Cummings B, McLeod R, Zuraw L, Zwaal C (2003) Follow-up of patients with curatively resected colorectal cancer: a practice guideline. BMC Cancer 3(1):26

[8] Graham RA, Wang S, Catalano PJ, Haller DG (1998) Postsurgical surveillance of colon cancer: preliminary cost analysis of physician examination, carcinoembryonic antigen testing, chest x-ray, and colonoscopy. Ann Surg 228(1):59

[9] Grossmann EM, Johnson FE, Virgo KS, Longo WE, Fossati R (2004) Follow-up of colorectal cancer patients after resection with curative intent–the GILDA trial. Surg Oncol 13(2):119–124

[10] Grossmann I, de Bock G, van de Velde C, Kievit J, Wiggers T (2007) Results of a national survey among Dutch surgeons treating patients with colorectal carcinoma. Current opinion about follow-up, treatment of metastasis, and reasons to revise follow-up practice. Colorectal Disease 9(9):787–792

[11] Grossmann I, Avenarius JK, Mastboom WJ, Klaase JM (2010) Preoperative staging with chest CT in patients with colorectal carcinoma: not as a routine procedure. Ann Surg Oncol 17(8):2045–2050

[12] Grossmann I, Verberne C, de Bock G, Havenga K, Kema I, Klaase J, Renehan A, Wiggers T (2011) The role of high frequency dynamic threshold (HiDT) serum carcinoembryonic antigen (CEA) measurements in colorectal cancer surveillance: a (revisited) hypothesis paper. Cancers 3(2):2302–2315

[13] Grossmann I, Doornbos P, Klaase J, de Bock G, Wiggers T (2014) Changing patterns of recurrent disease in colorectal cancer. Eur J Surg Oncol 40(2):234–239

[14] de Haas RJ, Wicherts DA, Andreani P, Pascal G, Saliba F, Ichai P, Adam R, Castaing D, Azoulay D (2011) Impact

of expanding criteria for resectability of colorectal metastases on short-and long-term outcomes after hepatic resection. Ann Surg 253(6):1069–1079

[15] Hahnloser D, Nelson H, Gunderson LL, Hassan I, Haddock MG, O'Connell MJ, Cha S, Sargent DJ, Horgan A (2003) Curative potential of multimodality therapy for locally recurrent rectal cancer. Ann Surg 237(4):502

[16] Hemming K, Haines T, Chilton P, Girling A, Lilford R (2015) The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 350:h391

[17] Hemming K, Lilford R, Girling AJ (2015) Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Stat Med 34(2):181–196

[18] van Houwelingen J (1998) Roaming through methodology. III. Randomization at the level of the physicians. Ned Tijdschr Geneeskd 142(29):1662–1665

[19] Jeffery M, Hickey BE, Hider PN, et al (2007) Follow-up strategies for patients treated for non-metastatic colorectal cancer. Cochrane Database Syst Rev 1(1)

[20] de Jong K (2007) Multimodality treatment of liver metastases increases suitability for surgical treatment. Alimentary pharmacology & therapeutics 26(s2):161–169

[21] Kievit J (2002) Follow-up of patients with colorectal cancer: numbers needed to test and treat. Eur J Cancer 38(7):986–999

[22] Kobayashi H, Mochizuki H, Sugihara K, Morita T, Kotake K, Teramoto T, Kameoka S, Saito Y, Takahashi K, Hase K, et al (2007) Characteristics of recurrence and surveillance tools after curative resection for colorectal cancer: a multicenter study. Surgery 141(1):67–75

[23] Locker GY, Hamilton S, Harris J, Jessup JM, Kemeny N, Macdonald JS, Somerfield MR, Hayes DF, Bast Jr RC (2006) ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. J Clin Oncol 24(33):5313–5327

[24] Pfannschmidt J, Dienemann H, Hoffmann H (2007) Surgical resection of pulmonary metastases from colorectal cancer: a systematic review of published series. The Annals of thoracic surgery 84(1):324–338

[25] Pita-Fernandez S, Alhayek-Ai M, Gonzalez-Martin C, López-Calviño B, Seoane-Pillado T, Pértega-Díaz S (2014) Intensive follow-up strategies improve outcomes in nonmetastatic colorectal cancer patients after curative surgery: a systematic review and meta-analysis. Ann Oncol 26(4):644–656

[26] Primrose JN, Perera R, Gray A, Rose P, Fuller A, Corkhill A, George S, Mant D (2014) Effect of 3 to 5 years of scheduled CEA and CT follow-up to detect recurrence of colorectal cancer: the FACS randomized clinical trial. JAMA 311(3):263–270

[27] Renehan AG, Egger M, Saunders MP, T O'Dwyer S (2002) Impact on survival of intensive follow up after curative resection for colorectal cancer: systematic review and meta-analysis of

randomised trials. BMJ 324(7341):813

[28] Rodríguez-Moranta F, Saló J, Arcusa À, Boadas J, et al (2006) Postoperative surveillance in patients with colorectal cancer who have undergone curative resection: a prospective, multicenter, randomized, controlled trial. J Clin Oncol 24(3):386–393

[29] Rosen M, Chan L, Beart RW, Vukasin P, Anthone G (1998) Follow-up of colorectal cancer. Diseases of the colon & rectum 41(9):1116–1126

[30] Staab HJ, Anderer FA, Stumpf E, Fischer R (1978) Slope analysis of the postoperative CEA time course and its possible application as an aid in diagnosis of disease progression in gastrointestinal cancer. The American Journal of Surgery 136(3):322–327

[31] Tjandra JJ, Chan MK (2007) Follow-up after curative resection of colorectal cancer: a meta-analysis. Diseases of the colon & rectum 50(11):1783–1799

[32] Treasure T, Monson K, Fiorentino F, Russell C (2014) The CEA Second-Look Trial: a randomised controlled trial of carcinoembryonic antigen prompted reoperation for recurrent colorectal cancer. BMJ open 4(5):e004,385

[33] van de Velde CJ, Aristei C, Boelens PG, et al (2013) EURECCA colorectal: multidisciplinary mission statement on better care for patients with colon and rectal cancer in Europe. Eur J Cancer 49(13):2784–2790

[34] Verberne C, Zhan Z, van den Heuvel E, Grossmann I, Doornbos P, Havenga K, Manusama E, Klaase J, van der Mijle H, Lamme B, et al (2015) Intensified follow-up in colorectal cancer patients using frequent Carcino-Embryonic Antigen (CEA) measurements and CEA-triggered imaging: Results of the randomized 'CEAwatch' trial. Eur J Surg Oncol 41(9):1188–1196

[35] Verberne CJ, Nijboer CH, de Bock GH, Grossmann I, Wiggers T, Havenga K (2012) Evaluation of the use of decision-support software in carcino-embryonic antigen (CEA)-based follow-up of patients with colorectal cancer. BMC Med Inform Decis Mak 12(1):14, DOI 10.1186/1472-6947-12-14

[36] Wille-Jørgensen P, Laurberg S, Påhlman L, et al (2009) An interim analysis of recruitment to the COLOFOL trial. Colorectal Disease 11(7):756–758

[37] Wong SL, Mangu PB, Choti MA, Crocenzi TS, Dodd III GD, Dorfman GS, Eng C, Fong Y, Giusti AF, Lu D, et al (2009) American Society of Clinical Oncology 2009 clinical evidence review on radiofrequency ablation of hepatic metastases from colorectal cancer. J Clin Oncol 28(3):493–508

[38] Yamamoto M, Maehara Y, Sakaguchi Y, Mine H, Yamanaka T, Korenaga D, Okamura T (2004) Distributions in CEA doubling time differ in patients with recurrent colorectal carcinomas. Hepato gastroenterology 51(55):147–151

[39] Zhan Z, van den Heuvel ER, Doornbos PM, Burger H, Verberne CJ, Wiggers T, de Bock GH (2014) Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. J Clin Epidemiol 67(4):454–461

5

SURVIVAL ANALYSIS OF THE CEAWATCH MULTICENTRE CLUSTERED RANDOMIZED TRIAL

C. J. Verberne
Z. Zhan
E. R. van den Heuvel
F. Oppers
A. M. de Jong
I. Grossmann
J. M. Klaase
G. H. de Bock
T. Wiggers

This chapter has been published in British Journal of Surgery, Volume **104**, Issue 8, (2017) [20]. (C.J.V. and Z.Z. contributed equally to this work)

ABSTRACT

Background: The CEAwatch randomized trial showed that follow-up with intensive carcinoembryonic antigen (CEA) monitoring (CEAwatch protocol) was better than care as usual (CAU) for early postoperative detection of colorectal cancer recurrence. The aim of this study was to calculate overall survival (OS) and disease-specific survival (DSS).

Methods: For all patients with recurrence, OS and DSS were compared between patients detected by the CEAwatch protocol versus CAU, and by the method of detection of recurrence, using Cox regression models.

Results: Some 238 patients with recurrence were analysed (7.5 per cent); a total of 108 recurrences were detected by CEA blood test, 64 (55.2 per cent) within the CEAwatch protocol and 44 (41.9 per cent) in the CAU group (P = 0.007). Only 16 recurrences (13.8 per cent) were detected by patient self-report in the CEAwatch group, compared with 33 (31.4 per cent) in the CAU group. There was no significant improvement in either OS or DSS with the CEAwatch protocol compared with CAU: hazard ratio 0.73 (95 per cent 0.46 to 1.17) and 0.78 (0.48 to 1.28) respectively. There were no differences in survival when recurrence was detected by CT versus CEA measurement, but both of these methods yielded better survival outcomes than detection by patient self-report.

Conclusion: There was no direct survival benefit in favour of the intensive programme, but the CEAwatch protocol led to a higher proportion of recurrences being detected by CEA-based blood test and reduced the number detected by patient self-report. This is important because detection of recurrence by blood test was associated with significantly better survival than patient self-report, indirectly supporting use of the CEAwatch protocol.

5.1. INTRODUCTION

Patients with colorectal cancer who have undergone surgery with curative intent usually require follow-up because early detection of asymptomatic disease improves the probability of treatment success. Indeed, the options for cure in recurrent disease have increased over time, leading to higher cure rates in this patient group.[4] Although postoperative surveillance guidelines after colorectal cancer surgery differ between countries, there is consensus that imaging and carcinoembryonic antigen (CEA) measurements should be performed routinely. Current Dutch guidelines on follow-up recommend measurement of CEA levels, along with liver ultrasonography and chest X-rays every 3–6 months for the first 3 years (http://www.oncoline.nl/colorectaalcarcinoom).

Recent research has focused on whether intensifying current follow-up strategies could improve survival, and several studies have compared the survival benefits between different follow-up strategies.[13] Earlier studies showed a modest and clinically relevant gain in survival for intensive protocols [10], but more recent trials [14, 15] have not reported such benefits. As available treatments and imaging technologies for the detection of recurrent disease have improved, data from older studies are likely to be invalid in the current era.

The CEAwatch RCT [18] showed that intensive CEA-based postoperative screening after colorectal cancer surgery (CEAwatch protocol) could provide benefits in terms of earlier detection times and higher curative treatment rates compared with care as usual (CAU) following current Dutch guidelines. In this trial, the CEAwatch protocol consisted of CEA measurements every 2 months in the first 3 years, and once every 3 months during the fourth and fifth years, combined with annual imaging by CT.

The aim of the present analysis was to assess whether the shortened

detection time and increased curative treatment rate have been associated with an increase in overall (OS) and disease-specific (DSS) survival. A secondary objective was to investigate whether there has been a difference in survival in relation to the method of detection of recurrence, namely CT imaging, CEA-based blood test and patient self-report. The null hypothesis of the present study was that patients with recurrence detected by the CEAwatch protocol would have the same risk of OS and DSS as those with recurrence detected by the CAU protocol.

5.2. METHODS

The CEAwatch trial was a multicentre, stepped-wedge, cluster randomized trial in the Netherlands, which compared an intensive CEA-based follow-up protocol with the current national guideline (CAU). The trial had a unidirectional crossover design in which each cluster switched from control treatment to intervention treatment, with randomization used to allocate clusters to different switch moments. Eleven non-academic teaching hospitals were randomly grouped into five clusters. The trial started in October 2010, and every 3 months one cluster switched from the CAU follow-up protocol to the intensified follow-up protocol (Fig. 5.1). A detailed description of the trial design has been published elsewhere.[22] The CEAwatch trial (Netherlands Trial Register NTR2182) was approved by the Medical Ethics Committee of the University Medical Centre Groningen (METc-UMCG 2010.064) and by the local ethics committees of all participating centres.

5.2.1. PARTICIPANTS AND DATA COLLECTION

Patients were eligible for inclusion if they had primary colorectal cancer, AJCC stage I–III disease, and underwent R0 resection between 2007 and July 2012. Between October 2010 and July 2012, all patients who provided

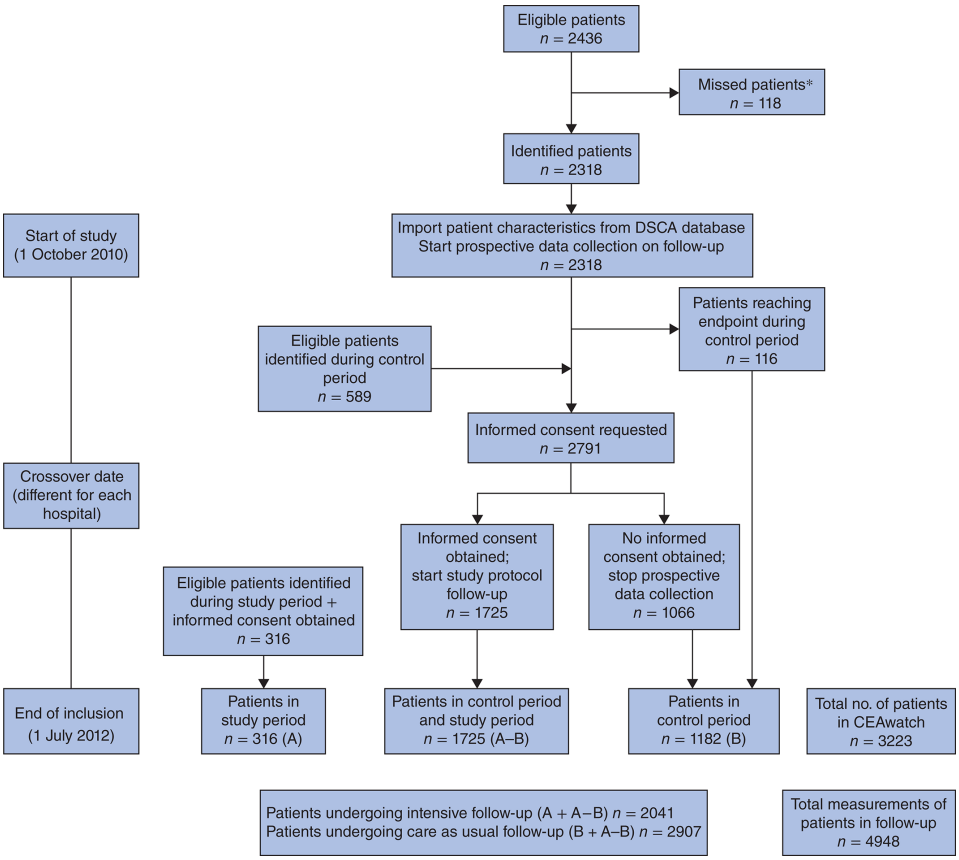

Figure 5.1 | Flow chart for patients with colorectal carcinoma in the CEAwatch trial. *Present in the Dutch Surgical Colorectal Audit (DSCA) database, but not found in hospital database

informed consent were assigned to the protocol of the hospital they were attending, and followed up until March 2015. For the present analysis, data on survival status (alive, dead), oncological status (recurrence, no recurrence), cause of death (cancer-related, not cancer-related), methods of detection of recurrence, and the treatment employed for recurrences were collected. The method of detection of recurrence was defined as that which indicated an abnormality (CEA, physical signs) leading to the imaging confirming recurrent disease. Data were updated until March

2015 by three investigators.

5.2.2. FOLLOW-UP PROTOCOLS

The control (CAU) follow-up protocol followed the national guidelines of the Netherlands in 2008. This comprised outpatient clinic visits every 6 months for the first 3 years and annual visits in years 4 and 5. CEA measurement was recommended every 3–6 months in the first 3 years and annually in following 2 years. It is known that adherence to this guideline is poor regarding the use of CEA measurement.[6] Liver ultrasonography and chest X-ray were recommended at each visit. Postoperative follow-up started after curative resection and adjuvant therapy, if this was given. Dutch oncology guidelines advise discussion of adjuvant chemotherapy with the patient in the case of AJCC stage III and II colonic cancer with a high risk of recurrence (small numbers of retrieved lymph nodes, perforated tumours, T4 N0 and other poor prognostic characteristics). Adjuvant treatment for rectal cancer is not advised in the Netherlands.

The intensive (CEAwatch) protocol consisted of taking CEA measurements every 2 months and performing annual imaging with high resolution CT of the thorax and abdomen for the first 3 years, followed by CEA measurement every 3 months in the fourth and fifth years. In this protocol, if the absolute CEA value was greater than 2.5 ng/ml and there was a 20 per cent increase in the measured CEA value from the previous one, another blood sample was drawn 4 weeks later.[7] If a rise in CEA level was noted between consecutive measurements, CT of the chest and abdomen was advised (Fig. 5.2). This intensive CEA protocol was based on a single-centre phase II trial, in which CEA levels were checked every month.[7] Based on these results, the authors were able to define the schedule with the highest sensitivity and specificity for detection of recurrence. Overall, the intensive CEAwatch protocol has more frequent CEA measurements and fewer clinical visits than the CAU protocol.

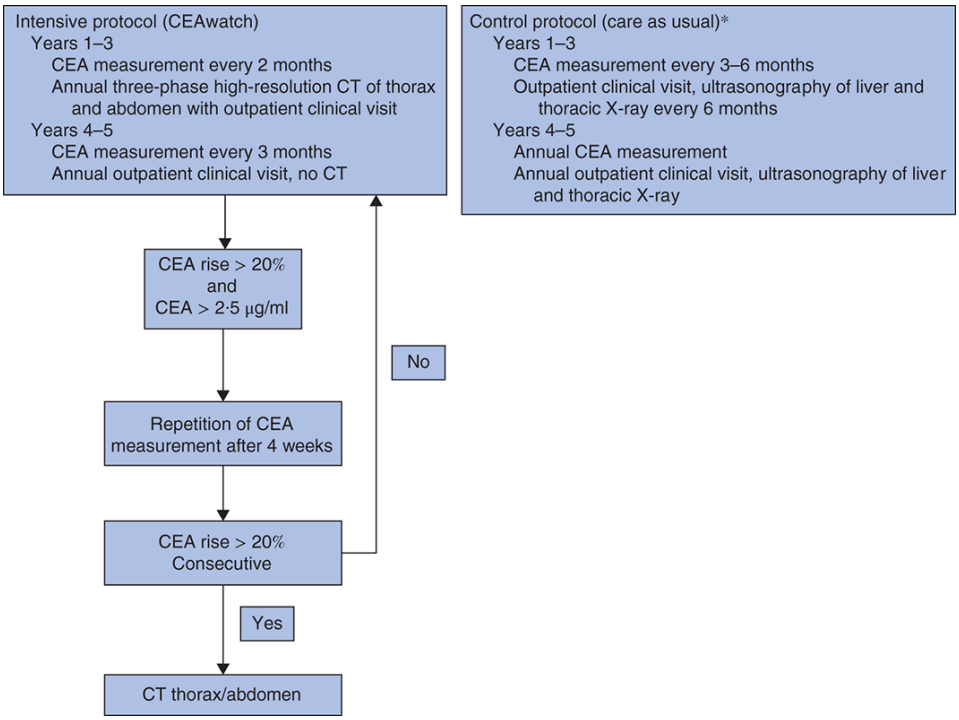

Figure 5.2 | Schematic illustration of the intensive carcinoembryonic antigen (CEA) measurement (CEAwatch) and control (care as usual) follow-up protocols. *Local differences and adjustment by individual hospitals allowed.

5.2.3. STATISTICAL ANALYSIS

Demographic and clinical characteristics of groups with and without recurrence during the study period were compared. Baseline data and recurrence detection methods were compared between the two follow-up protocols among patients with recurrence. The Mann–Whitney U test was used for analysis of continuous variables, and the χ^2 test for categorical variables or Fisher's exact test in case of insufficient cell count.

To investigate the effect of different follow-up strategies on survival, the analyses focused on patients with recurrence detected during the trial as the follow-up protocol is unlikely to influence survival among healthy (cured) patients. For both OS and DSS, a Cox Markov model [1, 12] was

used to compare the transition from recurrence to death among those with recurrence detected by the CEAwatch protocol versus the CAU protocol. The model was adjusted for age at diagnosis, sex, primary tumour stage and hospital. Hazard ratios (HRs) and 95 per cent confidence intervals were calculated.

Proportionality assumptions were verified by checking interactions between detection methods and logarithm of survival time, and by examining Schoenfeld residuals.[5, 17] When comparing the different detection methods, similar models were fitted by adjusting for the same co-variables and taking into account interactions between the detection methods and the follow-up protocol. If no significant effect was observed for the interaction, a model without these interactions was used in the final analysis.

For OS, survival curves were plotted for comparison both of the follow-up protocols and the detection methods. The survival functions were calculated by the Breslow estimator [3], taking into account the other co-variables instead of using the typical product-limit estimator of Kaplan–Meier curves. Adjusted survival curves are presented for the modal patient (typical patient with the most common characteristics) in the cohort.

For both OS and DSS, additional sensitivity analyses were performed to check the Markovian assumptions, that is whether the probability of death depended on the time of detection of a recurrence. For this, likelihood ratio tests were performed by including the time to detection as a co-variable in the Cox proportional hazard model.[12] In addition, an assessment was made of the assumption that the follow-up protocol had no effect on survival outcomes for patients without recurrence. This was done by fitting a Cox regression model to the data from the group without recurrence with follow-up protocol as a time-dependent co-variable. Two-sided P values are reported for all analyses; $P < 0.050$ was considered significant. All statistical analyses were conducted using SAS® 9.4

statistical software (SAS Institute, Cary, North Carolina, USA).

5.3. RESULTS

Between October 2010 and July 2012, a total of 3223 patients were included in the CEAwatch trial. Forty-one patients were excluded (based on inclusion/exclusion criteria during data analysis, 29; missing data on recurrence, 3; secondary tumour rather than recurrence, 9), leaving 3182 patients in the final analysis. A comparison of the baseline characteristics between patients with and without recurrence is shown in Table 5.1.

Table 5.1 | Baseline characteristics of patients and cancers by recurrence within the trial period

	Recurrence in trial period		P[†]
	No (n = 2944)	Yes (n = 238)	
Age at diagnosis (years)[*]	70 (63–77)	69 (62–78)	0.657[‡]
Sex ratio (M:F)	1630:1314	153:85	0.008
AJCC tumour stage			<0.001
I–II	2016 (68.5)	110 (46.2)	
III	928 (31.5)	128 (53.8)	
Primry tumour location			0.530
Colon (+ rectosigmoid)	1858 (63.1)	145 (60.9)	
Rectum	1086 (36.9)	93 (39.1)	
Adjuvant chemotherapy			0.005
No	2204 (74.9)	158 (66.4)	
Yes	740 (25.1)	80 (33.6)	
Patients with co-morbidity			0.493[§]
None	566 (45.1)	40 (40)	
Minor	593 (47.2)	53 (54)	
Major	97 (7.7)	6 (6)	
Unknown	1688	139	

Values in parentheses are percentages unless indicated otherwise;
[*] values are median (i.q.r.).
[†] χ^2 test, except
[‡] Mann–Whitney U test and
[§] Fisher's exact test.

In total, 238 patients (7.5 per cent) had recurrent disease detected

during the trial period; the median age at diagnosis of the primary tumour was 69 years. Of these, 153 (64.3 per cent) were men, 128 (53.8 per cent) had an AJCC stage III tumour, 145 (60.9 per cent) had primary tumours in the colon (including rectosigmoid), and 80 (33.6 per cent) underwent adjuvant chemotherapy. The baseline characteristics of patients with recurrence detected by the CEAwatch and CAU follow-up protocols are compared in Table 5.2. Apart from a significant difference in age at diagnosis, the main difference between the two groups was in the method by which recurrence was detected. The proportion of patients with recurrence detected by imaging was similar for both protocols, but a significantly higher proportion had recurrence detected by a CEA-based blood test rather than patient self-report in the CEAwatch compared with the CAU group.

5.3.1. OVERALL SURVIVAL

There was no significant difference in OS between patients with recurrence diagnosed within the CEAwatch protocol and those whose recurrence was detected with CAU (HR 0.73, 95 per cent c.i. 0.46 to 1.17) (Table 5.3 and Fig. 5.3). Patients who were older at diagnosis had a significantly higher risk of death than younger patients (HR 1.02, 1.00 to 1.04) and those with AJCC stage III disease had a higher risk of death than those with stage I or II (HR 1.48, 1.03 to 2.14).

For recurrences detected by the same method (CEA or imaging), there were no differences in survival between the two follow-up protocols (P = 0·496 for the overall interaction between detection method and follow-up protocol). There were also no statistically significant differences in the risk of death between patients whose recurrences were detected by CEA-based blood test and those detected by imaging (HR 1.34, 0.83 to 2.17) (Table 5.4 and Fig. 5.4). However, both CEA measurement (HR 0.39, 0.25 to 0.63) and imaging (HR 0.29, 0.17 to 0.51) were associated with a significantly lower

Table 5.2 | Baseline characteristics of patients with recurrence according to follow-up protocol

	Follow-up protocol		P[¶]
	Care as usual (n = 112)	CEAwatch (n = 126)	
Age at diagnosis (years)[*]	74 (64–80)	66 (61–74)	<0.001[#]
Sex ratio (M:F)	68:44	85:41	0.283
AJCC tumour stage			0.842
I–II	51 (45.5)	59 (46.8)	
III	61 (54.5)	67 (53.2)	
Primry tumour location			0.839
Colon (+ rectosigmoid)	69 (61.6)	76 (60.3)	
Rectum	43 (38.4)	50 (39.7)	
Detection method[†]			0.007[**]
Blood test (CEA)	44 (41.9)[‡]	64 (55.2)	
Imaging	28 (26.7)	36 (31.0)	
Patient self-report	33 (31.4)	16 (13.8)	

Values in parentheses are percentages unless indicated otherwise;
[*] values are median (i.q.r.).
[†] Information missing for 17 patients.
[‡] Three recurrences detected by both carcinoembryonic antigen (CEA) level and patient self-report, and one by both CEA level and imaging.
[§] One recurrence detected by both CEA level and patient self-report.
[¶] χ^2 test, except
[#] Mann–Whitney *U* test and
[**] Fisher's exact test.

Table 5.3 | Cox Markov model analysis to determine the effect of follow-up protocols on overall survival adjusted for patient characteristics

	Hazard ratio	P
Age (per year)	1.02 (1.00, 1.04)	0.026
Sex		
M	0.81 (0.56, 1.17)	0.261
F	1.00 (reference)	
AJCC tumour stage		
III	1.48 (1.03, 2.14)	0.035
I–II	1.00 (reference)	
Follow-up		
CEAwatch	0.73 (0.46, 1.17)	0.191
Care as usual	1.00 (reference)	

Values in parentheses are 95 per cent confidence intervals. The multivariable model was also adjusted for hospital. Check of dependency on time of detection of recurrence: parameter estimate -0.0005 (s.e. 0.0004); likelihood ratio 1.67, 1 d.f., P = 0.197.

risk of death than patient self-report (Table 5.4).

Table 5.4 | Comparison of the hazard ratio for overall survival between the different screening methods

	Reference group	Hazard ratio
CEA-based blood test	Imaging	1.34 (0.83, 2.17)
Imaging	Patient self-report	0.29 (0.17, 0.51)
CEA-based blood test	Patient self-report	0.39 (0.25, 0.63)

Values in parentheses are 95 per cent confidence intervals. CEA, carcinoembryonic antigen.

5.3.2. DISEASE-SPECIFIC SURVIVAL

The follow-up protocol (CEAwatch versus CAU) had no impact on the risk of death from colorectal cancer (HR 0.78, 95 per cent c.i. 0.48 to 1.28). However, older patients had a significantly higher risk of disease-

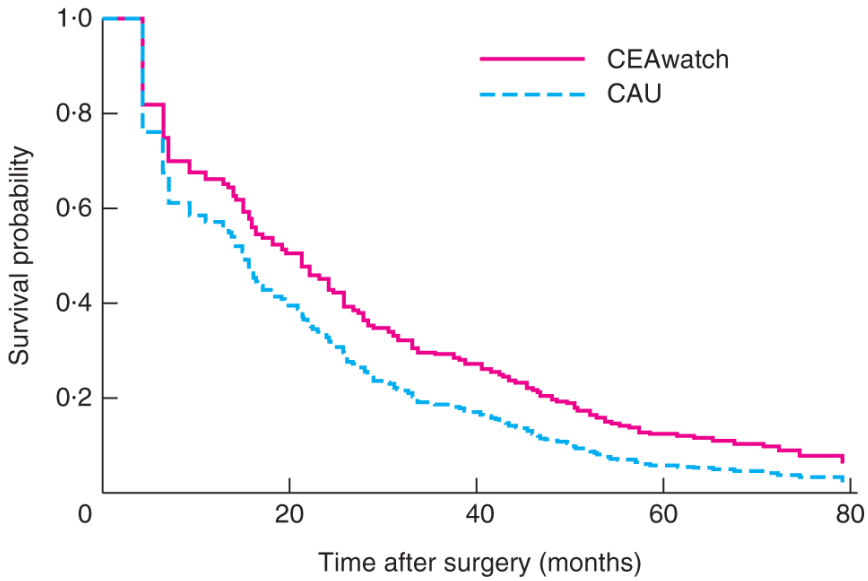

Figure 5.3 | Comparison of predicted overall survival for a typical modal patient with recurrence in two follow-up groups: intensive carcinoembryonic antigen measurement (CEAwatch) and care as usual (CAU)

specific death (HR 1.02, 1.00 to 1.04), as did patients with a higher tumour stage (AJCC III versus I–II) (HR 1.70, 1.14 to 2.52) (Table 5.5). As regards detection method, no differences in risk were observed between the CEA-based blood test and imaging methods (Table 5.6). In contrast, recurrence detected by CEA-based blood test (HR 0.33, 0.20 to 0.55) or imaging (HR of 0.26, 0.14 to 0.47) had a lower risk of colorectal cancer death than patient self-report.

5.3.3. SENSITIVITY ANALYSES

In verifying the assumption that the OS time did not differ between the two follow-up groups, the sensitivity analyses produced a non-significant HR of 0.75 (95 per cent c.i. 0.52 to 1.07), supporting the assumptions made in the main analyses.

Table 5.5 | Cox Markov model analysis to determine the effect of follow-up protocols on disease-specific survival adjusted for patient characteristics

	Hazard ratio	P
Age (per year)	1.02 (1.00, 1.04)	0.044
Sex		
M	0.78 (0.53, 1.16)	0.217
F	1.00 (reference)	
AJCC tumour stage		
III	1.70 (1.14, 2.52)	0.009
I–II	1.00 (reference)	
Follow-up		
CEAwatch	0.78 (0.48, 1.28)	0.328
Care as usual	1.00 (reference)	

Values in parentheses are 95 per cent confidence intervals. The multivariable model was also adjusted for hospital. Check of dependency on time of detection of recurrence: parameter estimate -0.0004 (s.e. 0.0004); likelihood ratio 1.29, 1 d.f., P = 0.255.

Table 5.6 | Comparison of the hazard ratio for disease-specific survival between the different screening methods

	Reference group	Hazard ratio
CEA-based blood test	Imaging	1.28 (0.76, 2.14)
Imaging	Patient self-report	0.26 (0.14, 0.47)
CEA-based blood test	Patient self-report	0.33 (0.20, 0.55)

Values in parentheses are 95 per cent confidence intervals. CEA, carcinoembryonic antigen.

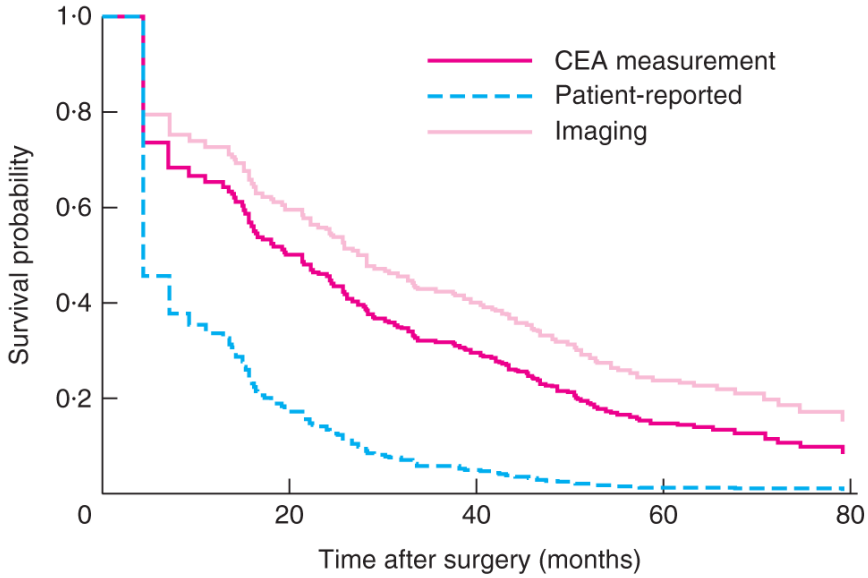

Figure 5.4 | Comparison of predicted overall survival for a typical modal patient with recurrence in relation to method of detection of recurrence. CEA, carcinoembryonic antigen

5.4. DISCUSSION

In this study, the intensive CEA-based follow-up protocol of the CEAwatch trial did not improve the OS or DSS of patients with recurrence. Similarly, the way in which recurrent disease was detected had no influence on either OS or DSS in comparisons of intensive CEA monitoring, conventional CEA measurements and imaging. However, OS and DSS were significantly worse when recurrences were detected by patient self-report alone compared with intensive CEA measurements (CEAwatch), conventional CEA measurements or imaging.

The findings of this study are in line with recent literature, including a meta-analysis.[13] In the FACS (Follow-up After Colorectal Surgery) trial [14], an RCT of 1202 patients with curatively resected colorectal cancer, participants were assigned to one of four follow-up groups: CEA alone

(300 patients), CT alone (299), CEA and CT (302) or minimal follow-up (301). The CEA level was considered raised to a degree that justified further investigation if it crossed an absolute threshold value of 7 ng/ml. In contrast to the CEAwatch trial, a dynamic threshold was not used, and the FACS trial reported a mean recurrence rate per year of 3.8 per cent after a mean 4.4 years of observation. Compared with minimal follow-up, intensive imaging or CEA screening increased the rates of surgical treatment with curative intent in patients with recurrence. As in the CEAwatch trial, intensive follow-up had no significant survival benefits over minimal follow-up, and all follow-up methods were more effective than no follow-up because of the poorer survival associated with waiting for recurrence to become symptomatic. The authors of the FACS trial suggested that the lack of survival gain with more intensive follow-up may have been due to a lack of study power to detect differences, and they intend to report again after a further 5 years of follow-up.

Similar conclusions were reached in the GILDA (Gruppo Italiano di Lavoro per la Diagnosi Anticipata) trial [15], an RCT comparing imaging follow-up strategies for colorectal cancer, in which 1228 patients were randomized to either minimal or intensive surveillance. Liver ultrasonography was performed twice over the 5-year study period in the minimal surveillance group, compared with every 4 months in the first year and often combined with chest X-rays in the intensive surveillance group. This contrasts with the CEAwatch trial, in which imaging was performed only once a year in the first 3 years. The mean recurrence rate per year was 4 per cent after a median follow-up of 5.2 years, which was higher than the 2.5 per cent per year observed in the CEAwatch trial. Despite this difference in imaging frequency, comparison of OS curves for the whole population showed no statistically significant differences. Although the COLOFOL study – a pragmatic randomized study to assess the frequency of surveillance tests after curative resection in patients with stage II and

III colorectal cancer – is now closed, no definitive results have been published. In the most recent report [2], published in January 2016, it was concluded that patients selected for COLOFOL were representative of the patient population suitable for follow-up.

With consistent evidence available from these three recent large-scale RCTs, the results of older studies, which suggested a clear survival gain with intensive follow-up, should no longer be considered relevant to the discussion of appropriate follow-up. These older studies are limited by more recent advances in diagnosis and therapy. Indeed, not only were they carried out before the routine use of preoperative CT, which has led to an increase in detection of synchronous metastases [8], they were also undertaken before critical advances were made in neoadjuvant and adjuvant treatments, as well as improvements in surgical techniques that have improved outcomes.[9, 11]

It has become clear from the FACS, GILDA and CEAwatch trials that both imaging and CEA measurement lead to the earlier detection of colorectal cancer recurrence during follow-up, but that regardless of the follow-up programme (imaging, CEA, or both), this is not directly translated into clear improvements in survival. However, a protocol with no follow-up was not included in any of the trials; even in the FACS study, a single CT scan was requested at study entry in the minimal follow-up arm. Given that there is sufficient evidence that treatment outcomes are worse for symptomatic than asymptomatic metastases, studies comparing follow-up with no follow-up after colorectal cancer treatment are unlikely to be justifiable. The present analysis confirmed this by showing significantly worse survival outcomes associated with recurrence detected by patient self-report than by CEA measurement or imaging. The CEAwatch protocol led to a higher proportion of recurrences being detected by CEA-based blood test and reduced the number detected by patient self-report. This might be because the intensive CEAwatch protocol has more frequent

CEA measurements and less frequent clinical visits than the CAU protocol. That the CEAwatch protocol led to a higher proportion of recurrences being detected by CEA-based blood test is important because those detected by blood test were associated with significantly better survival than those detected by patient self-report, indirectly supporting this intensive protocol. Thus, current evidence confirms the importance of postoperative follow-up, although the optimal protocol remains to be determined.

Colorectal cancer surveillance guidelines differ between countries, and empirical evidence to identify the best surveillance programme remains scanty [10]. The Dutch guideline still advises imaging with ultrasonography of the liver and thoracic X-rays, whereas most other countries recommend CT of the chest and abdomen, although this is not yet proven to be more effective. CEA testing is standard in all countries, but its frequency differs; most western European countries advise CEA tests between 3 and 6 months. Colonoscopic surveillance is recommended once every 5 years after surgery in both Great Britain and the Netherlands. A common recommendation is that follow-up should focus on patients with higher-stage primary tumours, as these are at greater risk of developing recurrence (for example, the European Society for Medical Oncology guideline [16]). However, this does not take into account the outcome of treatment of metastatic disease by primary tumour stage. Specifically, early recurrence after liver resection is associated with high primary tumour stage [21] and, as shown here, survival among patients with a lower primary tumour stage was slightly improved compared with that of patients with stage III tumours. This suggests that survival mainly depends on the biological characteristics of the primary tumour. Following this line of reasoning, it does not matter whether recurrent disease is detected and treated earlier in patients with high-stage primary tumours, because it is the biological tumour behaviour that ultimately lowers the survival probability, with earlier detection having little to no effect on the disease course. Thus, the

higher incidence of recurrences with poorer outcomes at high primary tumour stages must be balanced against the lower incidence of recurrences among low-stage tumours for which treatment is more effective.

The stepped-wedge trial design was statistically challenging. Owing to the sequential roll-out of the intervention, there were legitimate concerns that the majority of patients in the CEAwatch group would already have had a better prognosis because they had survived the CAU period, thereby leading to confounding. However, with the inclusion of patients who had undergone surgery before study enrolment (maximum of 3 years), and the dynamic recruitment during the trial, patient characteristics were considered to be balanced at different phases of the trial. As a result of the cluster design, this study was also prone to an unbalanced distribution of potential confounders. There were age differences between the two groups of patients (Table 5.2). Three centres, with an early switch time, had younger patients. Because of this susceptibility to an unbalanced distribution of potential confounders, these were corrected for in all the analyses.

The effect of the CEAwatch protocol on the transition from being disease-free after surgery (healthy state) to having recurrence detected (illness state) has been described previously by the CEAwatch trial investigators [18]. The present study focused on the transition from detection of recurrence to death. This approach is justifiable under the assumption that the underlying process is Markovian; that is, the probability of death depends only on the patient's current status.[1, 12] The sensitivity analysis showed that the data were consistent with the assumptions made during the analysis. It is also noteworthy that lead time bias is not of concern in the present analysis, as the full survival time, which was defined as the interval between the date of primary surgery and the event date, was taken into consideration.

The authors propose that quality of life and costs, rather than sim-

ply the primary tumour stage, should be considered when developing an optimal and efficient scheme for the follow-up of colorectal cancer. The motivation for the CEAwatch study was to find a more cost-effective follow-up system, based on the less costly CEA test, which could reduce the number of hospital visits during follow-up by identifying patients who need more than annual imaging. Although this strategy was cost-effective [19], the DSS and OS were comparable.

This study demonstrated no direct survival gain from the intensive CEAwatch follow-up programme. However, this protocol led to a higher proportion of recurrences being detected by a CEA-based blood test and a reduction in the number of recurrences detected by patient self-report. This is an important finding as survival is significantly improved if recurrence is detected while still asymptomatic. The data thus provide indirect support for postoperative follow-up with the new CEAwatch protocol. The authors propose that routine follow-up should now include CEA measurements to identify patients who need more frequent imaging.

ACKNOWLEDGEMENTS

C.J.V. and Z.Z. contributed equally to this work. The authors thank the following study coordinators of the participating hospitals in the CEAwatch trial: E. Manusama, H. van der Mijle, B. Lamme, K. Bosscha, P. Baas, B. van Ooijen, G. A. P. Nieuwenhuijzen, A. Marinelli, E. van der Zaag and D. Wasowicz. They also thank R. Sykes (www.doctored.org.uk) for editorial services. This study was funded by the Netherlands Organization for Health Research and Development (project numbers 171002209 and 171002211). The work of Z.Z. is supported by the Chinese Scholarship Council.

REFERENCES

[1] Andersen PK, Keiding N (2002) Multistate models for event history analysis. Stat Methods Med Res 11(2):91–115

[2] Andersson PH, Wille-Jørgensen P, Horváth-Puhó E, Petersen SH, Martling A, Sørensen HT, Syk I (2016) The COLOFOL trial: study design and comparison of the study population with the source cancer population. Clinical epidemiology 8:15

[3] Breslow NE (1972) Discussion of Professor Cox's paper. J Royal Stat Soc B 34:216–217

[4] Carpizo DR, D'angelica M (2009) Liver resection for metastatic colorectal cancer in the presence of extrahepatic disease. The lancet oncology 10(8):801–809

[5] Grambsch PM, Therneau TM (1994) Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81(3):515–526

[6] Grossmann I, de Bock G, van de Velde C, Kievit J, Wiggers T (2007) Results of a national survey among Dutch surgeons treating patients with colorectal carcinoma. Current opinion about follow-up, treatment of metastasis, and reasons to revise follow-up practice. Colorectal Disease 9(9):787–792

[7] Grossmann I, Verberne C, de Bock G, Havenga K, Kema I, Klaase J, Renehan A, Wiggers T (2011) The role of high frequency dynamic threshold (HiDT) serum carcinoembryonic antigen (CEA) measurements in colorectal cancer surveillance: a (revisited) hypothesis paper. Cancers 3(2):2302–2315

[8] Grossmann I, Doornbos P, Klaase J, de Bock G, Wiggers T (2014) Changing patterns of recurrent disease in colorectal cancer. Eur J Surg Oncol 40(2):234–239

[9] Hohenberger W, Weber K, Matzel K, Papadopoulos T, Merkel S (2009) Standardized surgery for colonic cancer: complete mesocolic excision and central ligation–technical notes and outcome. Colorectal disease 11(4):354–364

[10] Jeffery M, Hickey BE, Hider PN, et al (2007) Follow-up strategies for patients treated for non-metastatic colorectal cancer. Cochrane Database Syst Rev 1(1)

[11] MacFarlane J, Ryall R, Heald R (1993) Mesorectal excision for rectal cancer. The lancet 341(8843):457–460

[12] Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen PK (2009) Multi-state models for the analysis of time-to-event data. Stat Methods Med Res 18(2):195–222

[13] Mokhles S, Macbeth F, Farewell V, Fiorentino F, Williams N, Younes R, Takkenberg J, Treasure T (2016) Meta-analysis of colorectal cancer follow-up after potentially curative resection. Br J Surg 103(10):1259–1268

[14] Primrose JN, Perera R, Gray A, Rose P, Fuller A, Corkhill A, George S, Mant D (2014) Effect of 3 to 5 years of scheduled CEA and CT follow-up to detect recurrence of colorectal cancer: the FACS randomized clinical trial. JAMA 311(3):263–270

[15] Rosati G, Ambrosini G, Barni S, , et al (2015) A randomized trial of intensive versus minimal surveillance of patients with resected Dukes B2-

C colorectal carcinoma. Ann Oncol 27(2):274–280

[16] Schmoll H, Van Cutsem E, Stein A, Valentini V, Glimelius B, Haustermans K, Nordlinger B, Van de Velde C, Balmana J, Regula J, et al (2012) ESMO Consensus Guidelines for management of patients with colon and rectal cancer. a personalized approach to clinical decision making. Ann Oncol 23(10):2479–2516

[17] Schoenfeld D (1982) Partial residuals for the proportional hazards regression model. Biometrika 69(1):239–241

[18] Verberne C, Zhan Z, van den Heuvel E, Grossmann I, Doornbos P, Havenga K, Manusama E, Klaase J, van der Mijle H, Lamme B, et al (2015) Intensified follow-up in colorectal cancer patients using frequent Carcino-Embryonic Antigen (CEA) measurements and CEA-triggered imaging: Results of the randomized 'CEAwatch' trial. Eur J Surg Oncol 41(9):1188–1196

[19] Verberne C, Wiggers T, Grossmann I, Bock G, Vermeulen K (2016) Cost-effectiveness of a carcinoembryonic antigen (CEA) based follow-up programme for colorectal cancer (the CEA Watch trial). Colorectal Disease 18(3)

[20] Verberne CJ, Zhan Z, van den Heuvel ER, Oppers F, de Jong AM, Grossmann I, Klaase JM, de Bock GH, Wiggers T (2017) Survival analysis of the CEAwatch multicentre clustered randomized trial. Br J Surg 104(8):1069–1077, DOI 10.1002/bjs.10535, URL http://dx.doi.org/10.1002/bjs.10535

[21] Viganò L, Capussotti L, Réal Lapointe MD F, Barroso E, Hubert C, Giuliante F, Jan N, Mirza DF (2014) Early recurrence after liver resection for colorectal metastases: risk factors, prognosis, and treatment. a LiverMetSurvey-based study of 6,025 patients. Ann Surg Oncol 21(4):1276

[22] Zhan Z, van den Heuvel ER, Doornbos PM, Burger H, Verberne CJ, Wiggers T, de Bock GH (2014) Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. J Clin Epidemiol 67(4):454–461

6

PSYCHOLOGICAL EFFECTS OF THE INTENSIFIED FOLLOW-UP OF THE CEAWATCH TRIAL AFTER TREATMENT FOR COLORECTAL CANCER

Z. Zhan
C.J. Verberne
E. R. van den Heuvel
I. Grossmann
A. V. Ranchor
T. Wiggers
G. H. de Bock

This chapter has been published in PLOS One, Volume **12**, Issue 9, (2017) [23].

ABSTRACT

Background: The aim of the study was to evaluate psychological effects of the state-of-art intensified follow-up protocol for colorectal cancer patients in the CEAwatch trial.

Methods: At two time points during the CEAwatch trial questionnaires regarding patients' attitude towards follow-up, patients' psychological functioning and patients' experiences and expectations were sent to participants by post. Linear mixed models were fitted to assess the influences and secular trends of the intensified follow-up on patients' attitude towards follow-up and psychological functioning. As secondary outcome, odds ratios were calculated using ordinal logistic mixed model to compare patients' experiences to their expectations, as well as their experiences at two different time points.

Results: No statistical significant effects of the intensified follow-up were found on patients' attitude towards the follow-up and psychological functioning variables. Patients had high expectations of the intensified follow-up and their experiences at the second time point were more positive compared to the scores at the first time point.

Conclusion: The intensified follow-up protocol posed no adverse effects on patients' attitude towards follow-up and psychological functioning. In general, patients were more nervous and anxious at the start of the new follow-up protocol, had high expectations of the new follow-up protocol and were troubled by the nuisances of the blood sample testing. As they spent more time in the follow-up and became more adapted to it, the nervousness and anxiety decreased and the preference for the frequent blood test became high in replacement of conversations with the doctors.

6.1. INTRODUCTION

Recent studies investigating follow-up strategies for colorectal cancer (CRC) patients after treatment have provided favourable evidence for more intensive follow-up protocols using the measurement of serum carcinoembryonic antigen (CEA). It has been shown that intensive follow-up protocols are associated with higher detection rate of curative recurrences and shorter detection time compared to a minimal follow-up strategies or less intensive ones.[5, 13, 14, 18] In addition, ranging from non-significant to modest survival benefits have been reported by some studies as well.[9, 14, 15] Nowadays, such intense follow-up scheme has become guidelines for routine practice.[3, 12, 19]

The CEAwatch trial [19] is a multicentre randomized controlled trial conducted in the Netherlands between year 2010 and 2012. In this trial, the intensified follow-up protocol adheres bimonthly CEA measurements in the first three years and trimonthly CEA measurements during the fourth and fifth years combined with CT imaging. The control follow-up protocol is the Dutch care as usual follow-up guideline of which consists every 3-6 months CEA measurement and outpatient clinic visit every six months for the first three years and yearly CEA measurement and outpatient visit during the fourth and fifth year. Compared to the care as usual follow-up, the trial showed that the recurrences are detected earlier by the intensified follow-up protocol such that higher proportion of recurrences can be treated with curative intent.[19]

There is however no information with regards to the influences of the intensified follow-up protocol on the psychological aspects of patients and patients acceptance. Concerns have risen on the effects of high frequent CEA measurements and with that frequent reminders of the past disease, and the protocol that includes less frequent outpatient clinic visits and communication of test results by letters. From an implemen-

tation perspective, considering the medical benefits, the psychological outcomes should be at least comparable with the care as usual follow-up protocol.

The primary objective of the CEAwatch trial was to compare the CEA-watch follow-up scheme with the care as usual in terms of recurrence rate and detection time for the recurrences. Secondary outcomes considered were: quality of life, cost-effectiveness, and patients' survival. The aim of the here presented analysis was to evaluate the psychological effects of the intervention follow-up protocol in the CEAwatch trial, including the impact of more frequent blood sample testing on patients' psychological burden and worrisome of cancer, and explore patients' experiences and expectations of the new follow-up protocol. The null hypothesis was that the intensified follow-up has no effects on patients' attitude towards follow-up and psychological functioning. It was expected that a higher measurement frequency might on one hand give more burden and worries to patients and on the other hand might provide more reassurance. In addition, it was expected that patients would need time to adjust for the new follow-up protocol. The primary outcomes of this psychological evaluation study were patients' attitude towards the follow-up and their psychological functioning including anxiety and depression, fear of recurrences and cancer worries. The secondary outcomes were patients' experiences and expectations of the intensified follow-up.

6.2. MATERIALS AND METHODS

6.2.1. STUDY DESIGN

The assessments of patients' psychological variables were performed alongside the CEAwatch trial (Netherlands Trial Register 2182, URL: `www.trialregister.nl`, Date Registered: 26-Jan-2010). A detailed description of the trial has previously been published.[19] The CEAwatch trial

is a multi-centre stepped wedge cluster randomized trial (SW-CRT) conducted between 1st October, 2010 and 1st October, 2012 with eleven participating hospitals from the Netherlands. Patients were recruited during the period of 1st October, 2010 and 1st July, 2012. The trial was approved by the Medical Ethics Committee of the University Medical Centre Groningen (METc-UMCG2010.064) on 31st May 2010 and signed local feasibility declaration were obtained from all the local participating centres (Medical Ethical/Testing Committee of the Martini Ziekenhuis Groningen, Medisch Centrum Leeuwarden, Nij Smellinghe Drachten, Medisch Spectrum Twente Enschede, Meander Medisch Centrum Amersfoort, Jeroen Bosch Ziekenhuis Den Bosch, Albert Schweitzer Ziekenhuis Dordrecht, Medisch Centrum Haaglanden Den Haag, Gelre Ziekenhuis Apeldoorn, Catharina Ziekenhuis Eindhoven, and Elisabeth Ziekenhuis Tilburg). The authors confirm that all ongoing and related trials for this drug/intervention have been registered.

SW-CRT is a unidirectional design that allows the intervention to roll-out sequentially for all clusters of hospitals at different time periods of the trial.[6, 7, 22] At the beginning of a SW-CRT trial, all clusters start under the control and each cluster switches sequentially to the intervention at prespecified moments. All clusters remain under the intervention after the switch. The main motivation for adopting the SW-CRT design in the CEAwatch trial was that the computer aiding system used in the CEAwatch trial required time to be implemented at each site and SW-CRT provided logistic convenience by the phased introduction of the intervention.[22]

In the CEAwatch trial, hospitals were randomly grouped into five clusters and all clusters started with the care as usual follow-up protocol. Every three months, one randomly selected cluster switched from care as usual to intensified follow-up protocol (see Table 6.1). Written informed consents were obtained before the switch as required by the Medical Ethical Committee. During the trial, patients with AJCC stage I – III CRC

Table 6.1 | Follow-up schedule over time, according to the stepped wedge cluster-randomized design. At day 1 of every three-monthly period a new cluster switches from the care as usual protocol (CAU) to the intensified follow-up protocol (CEA). Grey periods 1 and 2 represent the times questionnaires were sent (1st round September 2011, 2nd round June 2012)

Cluster	Oct, 2010	Jan, 2011	Apr, 2011	Jul, 2011		Oct, 2011	Jan, 2012	
1	CAU	CEA	CEA	CEA		CEA	CEA	
2	CAU	CAU	CEA	CEA	1	CEA	CEA	2
3	CAU	CAU	CAU	CEA		CEA	CEA	
4	CAU	CAU	CAU	CAU		CEA	CEA	
5	CAU	CAU	CAU	CAU		CAU	CEA	

after curative treatment were included. Patients who received adjuvant chemotherapy were eligible after cessation of the adjuvant therapy. CONSORT diagram of the CEAwatch trial is provided in Fig 6.1.

The intensified follow-up protocol used in the CEAwatch trial adhered to every two months CEA measurements in the first three years and every three months CEA measurements during the fourth and fifth years of the follow-up. Evaluation of the rise in CEA was performed and an additional blood sample was drawn in case of CEA rise above 20% compared to the latest value, with minimum lower threshold CEA value 2.5 ng/ml. Outpatient clinic visits with imaging of thorax and abdomen were performed annually during the first three years of the follow-up. Blood test results (CEA value) including a laboratory form for the next appointment were sent to patients by automatically generated letters from a computer supporting system.[20] The care as usual follow-up followed the recommendation in the national guidelines of the Netherlands. This includes an outpatient clinic visit every six months for the first three years and annual visit during the fourth and fifth year, liver ultrasound and chest X-ray at each clinic visit, CEA measurements every 3-6 months for the first three years and once a year measurements during the fourth and fifth year.

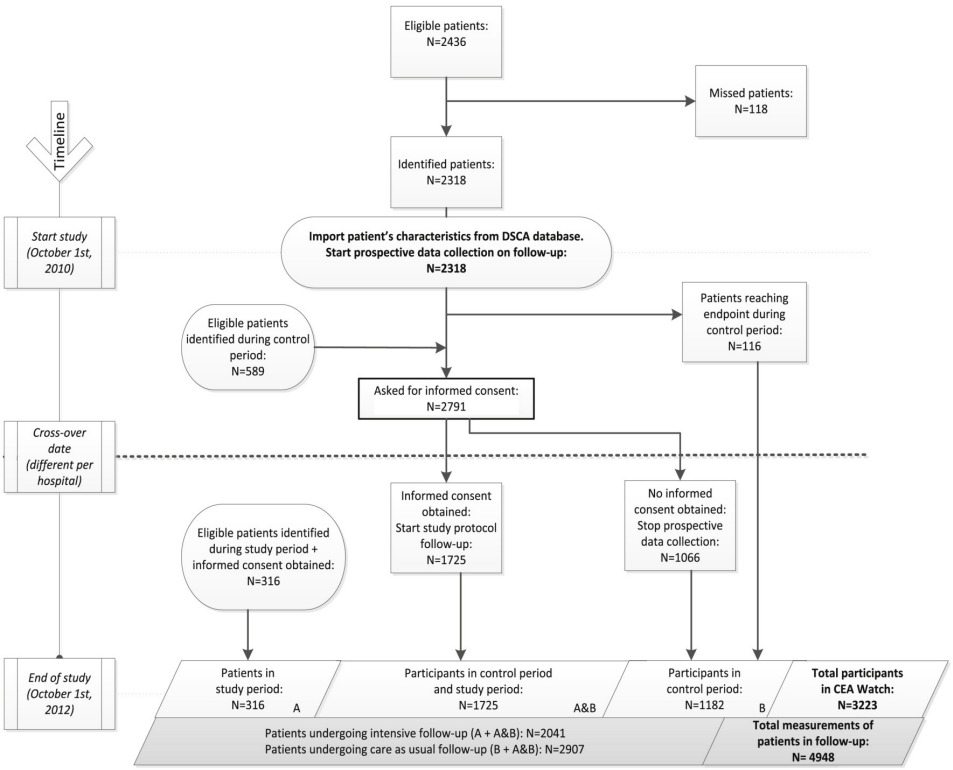

Figure 6.1 | Consort diagram of the CEAwatch trial.

6.2.2. DATA COLLECTION AND QUESTIONNAIRES

The psychological effects of the follow-up protocol were evaluated by questionnaires sent by post. As it was not permitted to collect data prior to the obtainment of the feasibility declaration from the local centre per requirement of the primary ethical committee, it was not possible to send out questionnaires while all clusters were exposed to the control follow-up protocol. Therefore, at two time points during the trial, patients were asked to fill in the questionnaires. The first time points was September 2011, after three of the five clusters (6 of the 11 hospitals) had already switched to the intensified follow-up and the other two clusters were still in the care as usual follow-up. The second time point was June 2012,

when all clusters had crossed over to the intensified follow-up and all patients had experienced the intensified follow-up (see Table 6.1). This had consequences of having different time between adopting intensified follow-up protocol and the psychological assessment. The durations of experiencing the new intensified follow-up protocols for patients from different clusters varied.

The questionnaires consisted of four sections: attitude towards follow-up, psychological functioning, experiences and expectations and sociodemographic data. Other disease-specific information, such as primary tumor stage, was retrieved from the CEAwatch trial.

ATTITUDE TOWARDS FOLLOW-UP

Patients' attitude towards the follow-up was measured by a validated 16-item questionnaire previously developed to assess routine follow-up of colorectal cancer.[17] The questionnaire consisted of four subscales: reassurance, nervous anticipation, perceived disadvantages of the follow-up and communication (with physicians). All items were measured with Likert scales ranging from 1 to 4. Items belonging to the same subscales were combined to derive a single sum scores for each subscale, respectively. For reassurance and communication, higher scores corresponded to more positive responses, while higher score corresponded to more negative responses for nervous anticipation and perceived disadvantages.

PSYCHOLOGICAL FUNCTIONING

The fear of recurrence was assessed by a 6-item questionnaire. From the original 3-item questionnaire used by several former studies [1, 17], this instrument was extended so that it is more tailored to the trial. The English translation of the added three items can be found in Table 6.2. Outcomes were measured with the sum scores of the 6 items ranging from 6 to 24. A higher score indicates stronger fear. The original 3-item questionnaire had a Cronbach's alpha of 0.75.[17] The extended version used in the

present study also had high reliability (Cronbach's alpha: 0.80) based on the current data. In addition, cancer worries were examined using the Dutch version of the validated Cancer Worry Scale [2, 4, 21], with each item using a 4-point Likert scale ranging from "never" to "almost always". General anxiety and depression were examined by the Dutch version of the Hospital Anxiety and Depression Scale (HADS).[16] It consisted of 14 items with 7 items for anxiety (ranging from 0 to 21) and 7 items for depression (ranging from 0 to 21). Within the HADS, a higher score meant more anxiety and depression respectively.

Table 6.2 | Extended questionnaires for the fear of recurrence.

	Item	Scale
Original	Do you feel insecure about your health?	Not at all – Very much
	Do you think the disease might still recur?	
	Do you feel completely cured?	
Extension	Do you feel that the disease will certainly come back to your bowel?	
	Are you afraid that the disease will come back somewhere else than the bowel?	
	If possible, would you prefer to go to a specialist nurses?	

EXPERIENCES AND EXPECTATIONS

For this part, a self-developed questionnaire was used. Patients were asked to complete 15 questions about their experiences during the intensified follow-up. If patients were still in the care as usual follow-up and had no experiences about the intensified follow-up, they were asked to answer the same 15 questions about the intensified follow-up to compare their expectations to the experiences. A 5-point Likert scale ranging from 1 to 5 was used for these items. These 15 questions are listed in Table 6.3.

Table 6.3 | Questionnaires regarding patients' experiences of the intensified follow-up protocol.

		⟵ More Positive		More Negative ⟶	
1) I am satisfied with the current follow-up	Totally agree	Agree	I don't know	Somewhat disagree	Completely disagree
2) I am afraid of blood tests[§]	Completely disagree	Somewhat disagree	I don't know	Agree	Totally agree
3) I find bimonthly blood tests[§]	Not stressful at all	Not stressful	I don't know	Somewhat stressful	Very stressful
4) Bimonthly check of my blood reassures me	Totally agree	Agree	I don't know	Somewhat disagree	Completely disagree
5) I would like my blood checked every two months	Totally agree	Agree	I don't know	Somewhat disagree	Completely disagree
6) Transportation for intensified follow-up is a problem for me[§]	Completely disagree	Somewhat disagree	I don't know	Agree	Totally agree
7) I hate to wait to turn in my blood sample[§]	Completely disagree	Somewhat disagree	I don't know	Agree	Totally agree
8) I find results send by letters very pleasant	Very pleasant	Pleasant	I don't know	Somewhat annoying	Very annoying
9) Knowing the dates of the blood testing results is of little importance to me[§]	Completely disagree	Somewhat disagree	I don't know	Agree	Totally agree
10) I think waiting a week for the blood test results is long[§]	Completely disagree	Somewhat disagree	I don't know	Agree	Totally agree
11) I think having a conversation with the doctor during visit is:	Very important	Important	I don't know	Somewhat unimportant	Completely unimportant
12) I think frequent testing for early detection of metastases is more important than a conversation with the doctor	Totally agree	Agree	I don't know	Somewhat disagree	Completely disagree
13) Having a conversation with the doctor once a year would be enough for me	Totally agree	Agree	I don't know	Somewhat disagree	Completely disagree
14) I would like to know if I have a metastasis , even though I'm aware this cannot be treated for months and I have no complaints	Totally agree	Agree	I don't know	Somewhat disagree	Completely disagree
15) I find it hard to cope with the uncertainty that the follow-up cannot guarantee the detection of the metastases[§]	Completely disagree	Somewhat disagree	I don't know	Agree	Totally agree

[§] The order of the options were deliberately reversed compared to the original questionnaire sent to patients so that OR>1 always indicates higher probability of being more positive.

6.2.3. STATISTICAL ANALYSES

The outcomes on the eight subscales, namely reassurance, nervous anticipations, disadvantages, communications, HADS anxiety, HADS, depression, cancer worry scores, and fear of recurrences, were considered as the primary outcomes. Patients' expectations and experiences were considered as secondary outcomes.

The aforementioned SW-CRT design required special attention of the secular trends in the analysis of the questionnaire data. Considering the nested structure of the design, a linear mixed model was used to assess the effects of the intensified follow-up on patients' attitude towards the follow-up and their psychological function corrected for the secular trends. Each primary outcome was considered separately as the dependent variable in the linear mixed model. To be more specific, for each dependent variable, three types of effects were assumed, namely the time effect, the treatment effect and the differences between patients who switched from control to intervention and those who experienced intervention only for both measurement rounds. Time effect was estimated by contrasting second time measurements to the first time measurements within the group of patients who only had intervention for both rounds. Differences between the two groups of patients were assessed by comparing the two groups at the second time point. The treatment effect was estimated by contrasting two treatment groups (intensified CEA compared to CaU) at the first time but correcting for the differences between patients. The psychological effects of the follow-up protocol were also corrected for age, gender and tumor stage. Outcomes from two measurement time points were modeled as bivariate normal and hospital was considered as a random effect. (Details about the linear mixed model can be found in S1.) The p-values of the hypothesis test were adjusted for multiplicity of testing several primary outcomes [11] using the Hochberg method.[8] Since patients' scores were not normally distributed within the attitude and psychological function-

ing dimensions, sensitivity analysis was conducted. These outcomes were reanalyzed with proper transformation of the outcome, namely logarithm and square root transformations. To keep the interpretation of the results simple and straightforward, the results of the linear mixed model were reported unless the sensitivity analysis would demonstrate a contradiction in conclusions. In that case, the results of the sensitivity analysis were reported instead.

To evaluate patients' experiences and expectations of the intensified follow-up, an ordinal logistical mixed model with cumulative logit link function was applied and odds ratios were calculated for two comparisons. The first comparison is between patients' experiences and their expectations corrected for the temporal effect. The second one is between patients' experiences measured at the 2nd time point and the experiences measured at the 1st time point. The model was also adjusted for patients' age, gender and the tumor stage. Factor analysis suggested no satisfying structural relationships among these 15 items by examining the Scree plot. Thus, the analysis was done item by item. No adjustment for multiple comparisons were made for this secondary outcome.[11] Only the odds ratios between experiences and expectations, as well as the odds ratios of experiences between the two time points, were presented in the result section.

If patients did not complete at least 80% of the items within certain subscales or dimensions, the score of this subscale/dimension was considered missing. Missing data was considered to be missing at random (MAR) and no special treatment for missing data was needed since inferences with maximum likelihood (used in both the linear and generalized mixed models) are still valid under this assumption. Statistical analyses were performed with SAS® statistical software, version 9.4. Linear mixed models were fitted using PROC MIXED and generalized linear mixed models were fitted using PROC GLIMMIX.

6.3. RESULTS

6.3.1. PATIENT CHARACTERISTICS AND RESPONSE RATE

On November 1st, 2011, total of 2,016 patients participated in the CEAwatch trial, and received the questionnaires. A total of 1,591 patients (78.9%) returned the questionnaires. On May 1st, 2012, total of 1,848 patients participated in the CEAwatch trial, 1556 (84.2%) of them returned the questionnaires. Patient characteristics of the two rounds are given in Table 6.4. During the first round, 820 (51.6%) of them participated in the care as usual follow-up and 770 (48.4%) were in the intensified follow-up (1 missing). At second round, all patients (2 missing) were in the intensified follow-up (Table 6.4). Among all patients, 1162 of them participated in both rounds of questionnaires. Summary of patients' experiences and expectations questionnaire is available in S1 Table.

6.3.2. PRIMARY OUTCOMES

The estimations for the psychological effects on patients' attitude towards follow-up and psychological functioning of the intensified follow-up protocol and time periods differences are shown in Table 6.5. No statistical significant effects of the intensified follow-up were found on patients' attitude towards the follow-up. Furthermore, there were no significant differences on anxiety and depression, fear of recurrences and cancer worries between the intensified follow-up protocol and care as usual follow-up. Comparing between two time points, no statistically significant temporal differences were found for all subscales.

6.3.3. SECONDARY OUTCOMES

The comparisons between patients' experiences and expectations are shown in Fig 6.2. In general, comparing patients' experiences in the

Table 6.4 | Patient characteristics and summary of primary outcome scores for the first round and second round evaluations.

	Round 1 (n=1591)	Round 2 (n= 1556)
Age: median (range)	68 (26-94)	68 (29-93)
AJCC stage[1]		
I	422 (27.80%)	433 (29.94%)
II	595 (39.20%)	572 (39.56%)
III	501 (33.00%)	441 (30.50%)
Gender[2]		
Female	685 (43.11%)	621 (40.01%)
Male	904 (56.89%)	931 (59.99%)
CEA follow-up[3]		
Intervention	770 (48.43%)	1554 (100.00%)
Control	820 (51.57%)	0 (0.00%)
Attitude towards follow-up	median (range)	median (range)
Reassurance	13 (4-16)	13 (4-16)
Nervous anticipation	7 (5-20)	7 (5-18)
Perceived disadvantages	4 (3-11)	4 (3-11)
Communication	13 (4-16)	13 (4-16)
Psychological functioning	median (range)	median (range)
Fear of recurrence	12 (6-24)	12 (6-22)
HADS: Anxiety	3 (0-21)	3 (0-21)
HADS: Depression	2 (0-20)	1 (0-20)
Cancer worries	13 (8-31)	13 (8-31)

[1] Missing 73 for round 1 and missing 110 for round 2;
[2] Missing 2 for round 1 and missing 4 for round 2;
[3] Missing 1 for round 1 and missing 2 for round2.

intensified follow-up to their expectations, the responses were towards the negative end of the spectrum. Particularly, patients expressed that the stress of the blood test was higher than they expected (OR: 0.10, 95% CL: [0.06, 0.16], p-value: <0.001) while they were less reassured by it (OR: 0.35, 95% CL: [0.24, 0.52], p-value: <0.001) and the preferences of the blood tests were not in favour of the intensified follow-up (OR: 0.22, 95% CL: [0.15, 0.33], p-value: <0.001). In addition, the inconveniences of the blood tests such as transportations (OR: 0.28, 95% CL: [0.14, 0.55], p-value: 0.0003), waiting time to turn in a blood sample (OR: 0.10, 95% CL: [0.06,

Table 6.5 | Estimates and 95% confidence limits of follow-up protocol effects and secular trends from linear mixed model for patients' attitude towards the follow-up and psychological functioning.

	Intensified follow-up vs. care as usual				Time trends			
	Estimates	95% CL		Adjusted p-value*	Estimates	95% CL		Adjusted p-value*
Reassurance	0.1202	-0.4504	0.6909	0.64	-0.2347	-0.5310	0.0617	0.42
Nervous anticipation	0.5738	-0.2669	1.4146	0.64	-0.5423	-0.9690	-0.1156	0.12
Disadvantage	0.2544	-0.2815	0.7904	0.64	-0.2153	-0.4880	0.0574	0.42
Communication	0.2365	-0.5618	1.0348	0.64	-0.3121	-0.7211	0.0967	0.42
HADS: Anxiety	0.6135	-0.0490	1.2759	0.56	-0.4348	-0.7925	-0.0771	0.12
HADS: Depression	0.3258	-0.4189	1.0706	0.64	-0.1461	-0.5319	0.2396	0.42
Cancer worries	0.2510	-0.7325	1.2346	0.64	-0.2275	-0.7319	0.2768	0.42
Fear of recurrence	0.2229	-0.8381	1.2838	0.64	-0.2264	-0.7651	0.3122	0.42

* Adjusted p-values were calculated according to the Hochberg method for multiple comparison adjustment.

0.18], p-value: <0.001) and results sent by letters (OR: 0.04, 95% CL: [0.02, 0.06], p-value: <0.001) were less appreciated.

In the comparisons between patients' second experiences and their first time experiences, the responses at the second time were more positive than the one at the first time as shown in Fig 6.3. At the second time points, patients had statistically significant higher probability to give a more positive response. Specifically, patients were more positive about all the items that did not meet with expectations in the previous comparison. Blood tests were less stressful (OR: 5.28, 95% CL: [3.91, 7.13], p-value: <0.001) and provided more reassurance (OR: 2.12, 95% CL: [1.66, 2.71], p-value: <0.001) at the second time point compared to their first time experiences. Preferences of the blood test became higher (OR: 2.75, 95% CL: [2.14, 3.52], p-value: <0.001) and the frequent tests were more preferred in replacement of having conversation with the doctors (OR: 1.89, 95% CL: [1.49, 2.41], p-value: <0.001). Satisfaction of yearly conversation with the doctors became higher as well (OR: 1.75, 95% CL: [1.40, 2.18]. p-value: <0.001) with the importance of the conversation with the doctors decreased (OR: 0.70, 95% CL: [0.52, 0.95]. p-value: 0.02). Furthermore,

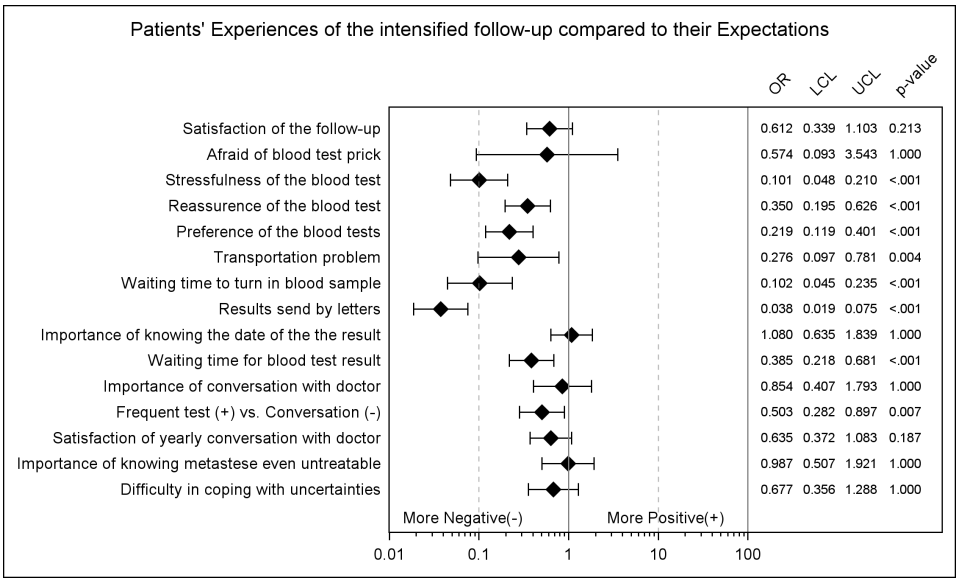

Figure 6.2 | Patients experiences of the intensified follow-up compared to their expectations.

6

patients felt easier coping with uncertainties of the test (OR: 1.32, 95% CL: [1.03, 1.71], p-value: 0.03)

6.3.4. SENSITIVITY ANALYSIS

The hypothesis tests of the linear mixed model could be affected by the skewed residual of the data. For reassurance subscale, the conditional residual was negatively skewed and the dependent variable itself was first converted to positive skewness and then logarithm-transformed. The estimations after the transformation (both treatment effect and time effect) were more towards the null and were consistent with the estimations of the linear mixed model. For nervous anticipation and cancer worry subscale, direct logarithm transformations were applied respectively. The treatment effect remained non-significant and the time effect remained significant for nervous anticipation. Both effects were shifted towards

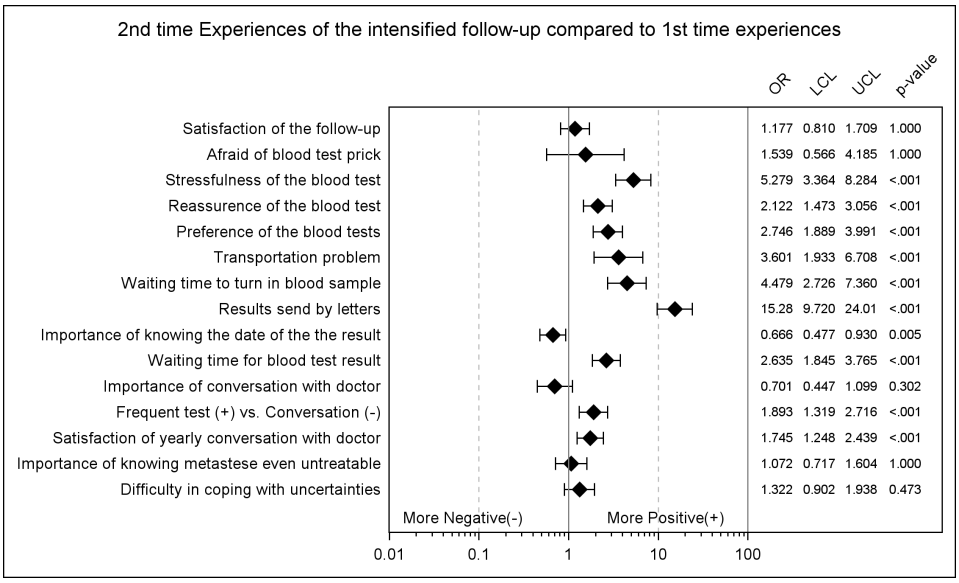

Figure 6.3 | Patients' 2nd time experiences of the intensified follow-up compared to their 1st time experiences.

the null for cancer worry subscale. For both HADS subscales, square root transformations were used and the results remained the same. The rest of the subscales were normally distributed. Detailed sensitivity analysis results are available in the S2 Table. To conclude, the results of the sensitivity analysis agreed with the linear mixed model and the estimations presented were accurate enough to be clinically meaningful.

6.4. DISCUSSION

In the CEAwatch trial, an intensified follow-up protocol was compared to the Dutch care as usual follow-up guideline. The major differences in the intensified follow-up protocol relevant to the discussion of the present study was that the frequency of outpatient clinic visit during the first three years of the follow-up were reduced and in replacement was a more intensive CEA measurements scheme.

The effects of the intensified follow-up protocol for CRC patients after surgery in the CEAwatch trial were evaluated with regards to patients' psychological variables. No statistical significant effects were found on patients' attitude towards the follow-up and psychological functioning. For patients' psychological functioning, no proof of increased burden or improvement was observed comparing the intensified follow-up protocol to the care as usual follow-up protocol.

Comparisons between patients' experiences and expectations resulted in more negative responses for patients' experiences which indicate that the expectations of the new follow-up protocols were high. On the other hand, by analysing the experiences at two different time points, we found that the responses became more positive later in time. Especially, patients responded more positively to blood test including reassurance, stressfulness and preference. This is in accordance with the results from the primary outcome that no decrease in reassurance were observed since it has been shown that patients are reassured by outpatient clinic visits and having conversation with the doctors.[17] From the present study, one may deduce that the frequent blood test compensated for less frequent clinic visit in the intensified follow-up protocol in terms of reassurance. In addition, patients' responses to the inconveniences of the blood tests were improved with time as well.

It has been mentioned that follow-up may remind patients of their cancers and possible relapsing of malignant disease.[17] However, even with more frequent blood tests, patients' cancer worries and fear of recurrences did not increase, nor did the HADS anxiety scores., Though it is expected that patients were more nervous and anxious about the new follow-up protocol as they were inexperienced with this new strategy, no significant differences were found between the earlier assessment time point and the later time point. On the other hand, from the exploratory analysis results of patients' experiences and expectations, it was indicated

that patients' preferences with the proposed intensified follow-up protocol increased as they became more familiar with the protocol.Currently, limited information is available regarding the impact of follow-up protocols on patients' quality of life and psychological functioning [10, 17] from the literature. The FACS study also planned to investigate the quality of life and satisfaction of care of the colorectal cancer follow-up and the results have not been published yet. The presented study with large sample size and high response rate, provided such information for the state-of-art post-treatment follow-up protocol. It should be noted that the statistical method used in the present study explicitly assumed no treatment-period interaction. In case treatment-period interaction was present, the estimator would be biased and could lead to an opposite conclusion on treatment effect. Thus our results should be viewed under the assumption of no treatment-period interaction effect. Due to the restricted policies from the medical ethical committee and the required time for collecting questionnaires by post, collecting data from more periods was infeasible. As such, it prohibits the possibility to investigate and verify the assumption of no treatment-period interaction which we originally planned for. Meanwhile, the results of the secondary outcome should be interpreted with caution since for this study relevant questions were formulated and these were analyzed item by item. The purpose was to provide a qualitative insight in patient's expectations and experience, tailored to the features of the intensified follow-up protocols used in the CEAwatch trial. In our opinion, it is sufficient enough to provide indirect evidence on the general trends of patients' experiences with regards to the intensified follow-up and is in agreement with the primary outcomes. In addition, doubts have been raised as to the validity of the HADS. It is recommended not to use this instrument anymore for future study. However, the questionnaires were already used by then.

In conclusion, the intensified follow-up protocol posed no adverse

effects on patients' attitude towards the follow-up and psychological functioning. In general, patients had high expectations of the new follow-up protocol and were troubled by the nuisances of the blood sample testing at the start of the new follow-up protocol. As they spent more time in the follow-up and became more adapted to it, the preference for the frequent blood test became high in replacement of conversations with the doctors.

REFERENCES

[1] de Bock G, Bonnema J, Zwaan R, van de Velde C, Kievit J, Stiggelbout A (2004) Patient's needs and preferences in routine follow-up after treatment for breast cancer. Br J Cancer 90(6):1144

[2] Custers JA, van den Berg SW, van Laarhoven HW, Bleiker EM, Gielissen MF, Prins JB (2014) The cancer worry scale: detecting fear of recurrence in breast cancer survivors. Cancer Nurs 37(1):E44–E50

[3] Duffy M, van Dalen A, Haglund C, Hansson L, Holinski-Feder E, Klapdor R, Lamerz R, Peltomaki P, Sturgeon C, Topolcan O (2007) Tumour markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines for clinical use. Eur J Cancer 43(9):1348–1360

[4] Eijzenga W, Aaronson NK, Hahn DE, Kluijt I, Ausems MG, Sidharta GN, Bleiker EM (2012) 19th annual conference of the international society for quality of life research. Qual Life Res 21(1):1–132, DOI 10.1007/s11136-012-0248-x, URL https://doi.org/10.1007/s11136-012-0248-x

[5] Figueredo A, Rumble RB, Maroun J, Earle CC, Cummings B, McLeod R, Zuraw L, Zwaal C (2003) Follow-up of patients with curatively resected colorectal cancer: a practice guideline. BMC Cancer 3(1):26

[6] Hemming K, Haines T, Chilton P, Girling A, Lilford R (2015) The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ 350:h391

[7] Hemming K, Lilford R, Girling AJ (2015) Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Stat Med 34(2):181–196

[8] Hochberg Y (1988) A sharper bonferroni procedure for multiple tests of significance. Biometrika 75(4):800–802

[9] Jeffery M, Hickey BE, Hider PN, et al (2007) Follow-up strategies for patients treated for non-metastatic colorectal cancer. Cochrane Database Syst Rev 1(1)

[10] Kjeldsen B, Thorsen H, Whalley D, Kronborg O (1999) Influence of follow-up on health-related quality of life af-

ter radical surgery for colorectal cancer. Scand J Gastroenterol 34(5):509–515

[11] Li G, Taljaard M, van den Heuvel ER, Levine MA, Cook DJ, Wells GA, Devereaux PJ, Thabane L (2016) An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int J Epidemiol 46(2):746–755

[12] Locker GY, Hamilton S, Harris J, Jessup JM, Kemeny N, Macdonald JS, Somerfield MR, Hayes DF, Bast Jr RC (2006) ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. J Clin Oncol 24(33):5313–5327

[13] Pita-Fernandez S, Alhayek-Ai M, Gonzalez-Martin C, López-Calviño B, Seoane-Pillado T, Pértega-Díaz S (2014) Intensive follow-up strategies improve outcomes in nonmetastatic colorectal cancer patients after curative surgery: a systematic review and meta-analysis. Ann Oncol 26(4):644–656

[14] Primrose JN, Perera R, Gray A, Rose P, Fuller A, Corkhill A, George S, Mant D (2014) Effect of 3 to 5 years of scheduled CEA and CT follow-up to detect recurrence of colorectal cancer: the FACS randomized clinical trial. JAMA 311(3):263–270

[15] Renehan AG, Egger M, Saunders MP, T O'Dwyer S (2002) Impact on survival of intensive follow up after curative resection for colorectal cancer: systematic review and meta-analysis of randomised trials. BMJ 324(7341):813

[16] Spinhoven P, Ormel J, Sloekers P, Kempen G, Speckens A, Van Hemert A (1997) A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of dutch subjects. Psychol Med 27(2):363–370

[17] Stiggelbout A, de Haes J, Vree R, van de Velde C, Bruijninckx C, van Groningen K, Kievit J (1997) Follow-up of colorectal cancer patients: quality of life and attitudes towards follow-up. Br J Cancer 75(6):914

[18] Tjandra JJ, Chan MK (2007) Follow-up after curative resection of colorectal cancer: a meta-analysis. Diseases of the colon & rectum 50(11):1783–1799

[19] Verberne C, Zhan Z, van den Heuvel E, Grossmann I, Doornbos P, Havenga K, Manusama E, Klaase J, van der Mijle H, Lamme B, et al (2015) Intensified follow-up in colorectal cancer patients using frequent Carcino-Embryonic Antigen (CEA) measurements and CEA-triggered imaging: Results of the randomized 'CEAwatch' trial. Eur J Surg Oncol 41(9):1188–1196

[20] Verberne CJ, Nijboer CH, de Bock GH, Grossmann I, Wiggers T, Havenga K (2012) Evaluation of the use of decision-support software in carcino-embryonic antigen (CEA)-based follow-up of patients with colorectal cancer. BMC Med Inform Decis Mak 12(1):14, DOI 10.1186/1472-6947-12-14

[21] Watson M, Duvivier V, Walsh MW, Ashley S, Davidson J, Papaikonomou M, Murday V, Sacks N, Eeles R (1998) Family history of breast cancer: what do women understand and recall about their genetic risk? J Med Genet 35(9):731–738

[22] Zhan Z, van den Heuvel ER, Doornbos PM, Burger H, Verberne CJ, Wiggers

T, de Bock GH (2014) Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. J Clin Epidemiol 67(4):454–461

[23] Zhan Z, Verberne CJ, van den Heuvel ER, Grossmann I, Ranchor AV, Wiggers T, de Bock GH (2017) Psychological effects of the intensified follow-up of the CEAwatch trial after treatment for colorectal cancer. PLOS ONE 12(9):e0184,740, DOI 10.1371/journal.pone.0184740, URL https://doi.org/10.1371/journal.pone.0184740

SUPPLEMENTARY MATERIAL

S1. STATISTICAL MODEL DESCRIPTIONS

The following model was fitted to each of the primary outcomes of the study separately:

$$Y_{ijk} = a_i + b_{k(i)} + \text{Age}_{ik} + \text{Gender}_{ik} + \text{AJCC}_{ik} + \text{Group}_{ij} + e_{ijk},$$

with Y_{ijk} as the outcome of kth patients at jth round from ith hopistal, a_i a normally distributed random effect of the hospital i, $b_{k(i)}$ a normally distributed random effect of patient k nested within hospital i, Age_{ik} a continuous variable representing patient's age when the measurement took place, Gender_{ik} a categorical variable for patient's gender (F=female, M=male), AJCC_{ik} a categorical variable for patient's AJCC tumour stage of the primary tumour (1=Stage I, 2=Stage II, 3=Stage III), Group_{ij} a categorical variable that was used for contrasting both the difference between the follow-up protocols and the difference between the two time points. For patients at the first time point experiencing the CAU follow-up protocol, the Group variable was coded as 1, and for patient at the first time point already exposed to the intensified CEA follow-up protocol, the Group variable was coded as 2. Furthermore, at the second time point, patients who were under CAU and under intensified CEA protocol in the previous round was coded as Group 3 and Group 4, respectively.

Treatment effects were estimated by the following contrast:

$$\text{Treatment Effect} = (\text{Group}\,3 - \text{Group}\,1) - (\text{Group}\,4 - \text{Group}\,2),$$

Period effects were estimated by the following contrast:

$$\text{Period Effect} = \text{Group}\,4 - \text{Group}\,2.$$

Group or cohort differences at follow-up were estimated with the contrasts:

$$\text{Cohort Effect} = \text{Group}\,4 - \text{Group}\,3.$$

Table S1 | Frequencies of the Likert scores for the secondary outcomes of patients' experiences and expectations of the intensified follow-up protocol measured at time point 1 and 2.

	Round 1					Round 2				
Likert Scale	1	2	3	4	5	1	2	3	4	5
Item (1)	60.52%	33.72%	1.92%	2.18%	1.66%	59.39%	34.40%	2.09%	2.94%	1.18%
Item (2)	4.16%	3.45%	2.50%	7.42%	82.47%	3.29%	3.16%	1.51%	7.31%	84.73%
Item (3)	2.05%	10.44%	2.82%	33.82%	50.86%	0.71%	5.78%	1.17%	32.42%	59.91%
Item (4)	40.31%	45.20%	8.63%	3.80%	2.06%	44.62%	45.14%	5.22%	3.59%	1.44%
Item (5)	41.91%	40.30%	8.77%	5.67%	3.35%	47.45%	40.72%	6.08%	3.40%	2.35%
Item (6)	3.23%	4.65%	3.56%	13.19%	75.37%	2.62%	2.43%	1.51%	7.74%	85.70%
Item (7)	5.09%	11.34%	3.48%	19.21%	60.89%	3.47%	9.31%	3.47%	16.12%	67.63%
Item (8)	27.98%	42.75%	12.89%	10.57%	5.80%	39.19%	50.46%	4.75%	4.49%	1.11%
Item (9)	25.45%	32.86%	11.56%	15.71%	14.42%	30.93%	35.85%	7.47%	14.88%	10.88%
Item (10)	13.61%	23.68%	11.55%	26.26%	24.90%	9.17%	21.04%	9.31%	26.47%	34.01%
Item (11)	52.84%	43.02%	2.39%	1.42%	0.32%	47.29%	45.24%	4.03%	2.58%	0.86%
Item (12)	44.19%	36.18%	10.59%	6.65%	2.39%	51.57%	32.37%	9.96%	4.78%	1.31%
Item (13)	15.40%	33.76%	10.63%	22.81%	17.40%	18.55%	37.37%	11.91%	19.01%	13.16%
Item (14)	51.68%	32.41%	11.28%	2.45%	2.19%	53.28%	31.63%	11.88%	1.57%	1.64%
Item (15)	13.96%	31.66%	17.37%	22.65%	14.35%	11.26%	30.71%	18.53%	24.10%	15.39%

Table S2 | Results of the sensitivity analysis

		Treatment effects					Time trends				
Y	Transform	Est.	95% CL		Raw p-value	Adjusted p-value	Est.	95% CL		Raw p-value	Adjusted p-value
Reassurance	$\log(17-Y)$	-0.0491	-0.2102	0.1120	0.5064	0.8185	0.0708	-0.0129	0.1546	0.0902	0.4833
Nervous anticipation	$\log(Y)$	0.0676	-0.0326	0.1679	0.1602	0.8185	-0.0662	-0.1172	-0.0153	0.0159	0.1272
Disadvantage	Y	0.2544	-0.2815	0.7904	0.3081	0.8185	-0.2153	-0.4880	0.0574	0.1091	0.4833
Communication	Y	0.2365	-0.5618	1.0348	0.5182	0.8185	-0.3121	-0.7211	0.0967	0.1210	0.4833
HADS: Anxiety	$\sqrt{Y}$	0.1340	-0.0543	1.2759	0.5600	0.8185	-0.4348	-0.7925	-0.0771	0.1200	0.2590
HADS: Depression	$\sqrt{Y}$	0.3258	-0.4189	0.3224	0.1630	0.8185	-0.1093	-0.2120	-0.0066	0.0370	0.4833
Cancer worries	$\log(Y)$	0.1005	-0.0805	0.2816	0.2763	0.8185	-0.0475	-0.1475	0.0524	0.3513	0.4833
Fear of recurrences	Y	0.2229	-0.8381	1.2838	0.6446	0.8185	-0.2264	-0.7651	0.3122	0.3711	0.4833

7

Discussion

7.1. SUMMARY

Stepped wedge design has become quite popular among epidemiologists and medical researchers for its flexibilities and efficiencies. [3, 4] As in reality, the benefits and drawbacks of the design are usually not as clear cut as in theory. Detailed and careful considerations are needed to properly evaluate the merits compared to other design options. [12] As a consequence, planning and conducting a stepped wedge design is challenging. There are relatively more factors to configure in the stepped wedge design compared to the classic parallel group design. Just to name a few, the number of measurement occasion, the number of switching moments, and the number of clusters per switch moment. All those factors play an important role in the development of a study with the stepped wedge design. Sample size calculations for a study with the stepped wedge design are complicated. Not only is the currently proposed sample size calculation method limited to normally distributed outcomes, there is also a lack of techniques to deal with some frequently encountered problems like unbalanced sample sizes, missing values and dropouts. In addition, it is not trivial to analyze data of a study with the stepped wedge design. [17] Not all standard methods for data analysis can be directly applied to data collected within the stepped wedge design and certain issues remained open. Therefore, the aim of the present thesis, motivated by a real life example of the CEA-Watch trial, was to investigate and demonstrate the practical issues of conducting a stepped wedge design such as the implications of applying a stepped wedge design for a study and the applications of frequently used statistical methods for a stepped wedge design.

In chapter 2, the strengths and weaknesses of the stepped wedge design were discussed and evaluated in the context of the specific example, namely the CEA-Watch trial. In the CEA-Watch trial, the stepped wedge design is beneficial in terms of increased sample size due to the motiva-

tion of participation to both patients and doctors. Combined with the common conveniences and flexibilities provided by the stepped wedge design, these features outweigh the complications caused by the informed consent and complexities of the statistical analysis. So it has been found that the majority of the benefits still hold true while some criticisms regarding the stepped wedge design are less problematic. For instance, it has become clear that the stepped wedge design does not necessarily prolong trial durations compared to other designs because of the required longitudinal setup in the CEA-Watch trial (multiple visits). The follow-up for eligible patients is 5 years. It is essential to have comparable trial duration (eg, 3–5 years) to investigate the effectiveness of the intensified follow-up routine no matter what kind of design is being used. Indeed, considerations should always be tailored to the specific trial at hand and general theoretical assumptions do not always hold true in practice. In general, the stepped wedge design is a reasonable alternative for parallel group design especially for large-scale pragmatic trials like the CEA-Watch trial.

Statistical analysis is difficult for the stepped wedge design for several reasons. First of all, the sequential introduction of treatment makes treatment a time-dependent variable and period a confounder for the treatment effect. In other words, period should be taking into account during the analysis. In Chapter 3, it has been demonstrated by simulations that ignoring period effects when there exist such effects would lead to biased estimations of the treatment effect and the estimated treatment effect no longer retains the interpretation of the average treatment effect. Secondly, since patients are solely exposed to one treatment during the first and the last period of a stepped wedge trial, period-treatment interactions would make the analysis even more difficult. In literature, the most frequently applied statistical method for the analysis of data in a stepped wedge trial is the linear mixed model presented by Hussey

and Hughes.[13] However, it assumes a constant treatment effect over the periods and the inclusion of treatment-period interactions is not as straightforward as in case of a parallel group design. This problem was demonstrated in Chapter 3 as well. Not all treatment period effects can be estimated, due to the set-up of the stepped wedge design. The overall treatment effect can be unbiasedly estimated only when we would know the pattern of treatment-period interactions (e.g., linear, random or constant). Thirdly, aggregated cluster-level analysis of the stepped wedge design is infrequently discussed in literature. In Chapter 3, we proposed to use meta-analysis techniques for the analysis of stepped wedge designs at a cluster level when period effects are not present. Meta-analysis techniques are strong candidates for the analysis since it provides the opportunity to assess treatment heterogeneities across clusters. Besides, theories are well-developed for meta-analysis to handle different types of outcomes.

In the CEA-Watch trial, we considered patients' disease progress as an illness-death process. [1] The healthy state in the illness-death process refers to the disease-free condition of patients. Once a recurrence was detected, the patient was considered to have transitioned into the illness state of the process. The absorbing death state of the process referred to the actual decease of the patients with or without recurrent metastasis. In chapter 4, the outcomes of interest were the probability of recurrence detection and the time-to-detection by the follow-up protocols. This chapter was primarily concerned with the effect of follow-up protocols on the transition probabilities from the disease-free state to the illness state. It should be mentioned that patients might already be deceased before the experience of recurrence. Therefore, death is considered as a competing risk event to the recurrence event. In the analysis of Chapter 4, we treated death without recurrence as a censored observation and considered a cause-specific hazard model. In chapter 5, the outcomes of

interest were the long term overall survival and disease-specific survival time of the patients. This chapter investigated the effect of the intensified follow-up protocol on the transition from illness state to the death state. These two chapters were essentially studies about the two transitions in the illness-death process. Alternatively, the analysis can be conducted using a more sophisticated multi-state model which incorporates both the illness-death process and the time-dependent treatment switching regime. Nonetheless, the presented analysis in Chapter 4 and 5 was sufficient for answering the clinical-relevant questions without hindering on the understanding of the findings for less technical readers.

Chapter 6 illustrated the analysis for a questionnaire type of data. The distinctive feature was that the outcomes were measured at only two different time points. At the first time point, some of the clusters had already switched to the new intervention while the other clusters were still in care as usual. At the second time point, all clusters had been switched. The approach shown was to parameterize the linear mixed model taking into account the structure of the design and contrasts the treatment effect in ad-hoc manner. It was essential to take into account the period effects in the analysis since it is very likely that patients' opinions about the follow-up protocol would change in time. Retrospectively, it would be preferable to have more measurements at different phases of the trial. However, considering the interval between each measurement and the time required to send and collect the questionnaire by post, it was considered not feasible to add more measurements on this point. Another implication of the limited number of measurement points is that investigations on the treatment-period interaction was not feasible. The proposed method explicitly assumes a constant treatment effect across periods, which unfortunately could not be verified.

7.2. PRACTICAL IMPLICATIONS

7.2.1. DESIGN CONSIDERATIONS

In general, three elements need to be considered in the designing phase of a stepped-wedge based trial. The first element is the type of the stepped wedge design to be used. There are two dimensions to be considered. The first one is the level of randomization. It is usually assumed in the literature that a stepped wedge design is a clustered randomized trial; however a stepped wedge design randomized at a patient level should not be ruled out. The second one is the type of cohort to be included in the trial. According to the characteristics of the target population and the disease of interests, the choice between a cross-sectional, longitudinal or open-cohort stepped wedge can be determined. The decision between the first two is frequently made based on the methods of measurements for the primary outcome. For instance, for a trial that aims to study the long term effectiveness of a drug for lowering patients' blood pressure, it would be illogical to use a cross-sectional stepped wedge design instead of a longitudinal one since it usually requires multiple measurements of the blood pressure to be able to provide sufficient evidence of any clinical-relevant effect. On the other hand, it would be impossible for a trial that studies the effect of a new surgical intervention technique on certain disease to repeat both the old and new procedure on the same patient, and therefore in that case, the cross-sectional stepped wedge design is the only and most obvious choice. It is sometimes less trivial but beneficial to also consider an open cohort stepped wedge design. In the CEA-Watch trial, an open cohort stepped wedge design was adopted. This design combines the longitudinal stepped wedge design with cross-sectional stepped wedge design. It is longitudinal due to the repeated measurement of the post-surgery follow-up protocol while new patients were included during each period making the design also cross-sectional. Due to the

pragmatic nature of the trial, which was to evaluate the effectiveness of the CEA-based follow-up protocol in real practice, it is important to include not only prevalent cases but also incident cases. Furthermore, an open cohort design provides a way to maintain a balanced sample size at each period and therefore protects the trial against issues caused by dropout and attrition. Since there is a lack of appropriate sample size calculation methods to adjust for dropout, the open cohort was more appealing at the design phase of the trial. Usually, studies of acute illness are of short period while chronic disease are of long term. Especially for a chronic disease such as cancer, attrition is deemed to be a critical factor to be accounted for. Meanwhile, an open cohort stepped wedge design is more cumbersome to implement, especially for large trials like CEA-Watch, to keep track of all patients and maintain database integrity. In the CEA-Watch trial, the medical ethic committee made the restriction such that no data can be collected until the cluster has switched to the intensified follow-up protocol. Thanks to the Dutch Surgical Colorectal Audit database, it was still possible to obtain data during the control period of the follow-up. In addition, an automated computer system [16] was deployed which had certainly ensured the quality of the trial to some extent.

The second element is the design-specific configurations. To be more specific, for stepped wedge design, three design parameters need to be considered, namely the number of measurements, the number of switches, and the number of clusters per switch. In practice, the number of measurement is dictated by the duration of the trial and the clinical requirement of the measurement frequency. The minimum trial duration should be considered to have sufficient follow-up period length and the measurement frequency is usually determined by the measurement method and logistic constraints. In a typical stepped wedge design, when a switch takes place at each period, the number of switch moments is one

less than the number of measurements. However, it is not necessary to switch at each period. Sometimes, it is not possible to do so since implementing the new intervention takes much longer time then one round of measurements. Furthermore, the maximum number of switch moments is limited by the number of clusters. For example, if there exits only 5 clusters but 20 measurements, there can be only a maximum of 5 switch moments, one switch per cluster. Therefore switches at each period may become impossible for certain trials. On the other hand, when there are more clusters, multiple clusters can have the same switch moment. For instance, in the CEA-Watch trial, the frequency of the measurement differs between the care as usual follow-up protocol and the CEA-Watch follow-up protocol. During the control period, patients' blood was sampled every 3-6 months while in the intervention period it was sampled every 2 months. Given the two-year trial period combined with 11 hospitals, it is not possible to switch every 2 months at hospital level. Therefore, the 11 hospitals were allocated to 5 switch moments.(Figure 1.1) To ensure a balanced sample size for each group, three smaller hospitals were grouped together.

The third element is the sample size and power calculations. Sample size and power calculation for a stepped wedge design is more complex than the classic parallel group design and cluster randomized designs. Not only the power of a trial is intimately related to the statistical methods and models which are not trivial even for normally distributed outcomes as shown in Chapter 3, but also the current existing tools are not flexible enough to handle many real life challenges such as unbalanced clusters and dropouts. Nevertheless, a blind application of a calculation formula in the literature will inevitably lead to inadequate sample sizes. At the time being, simulation based sample size calculation is preferred [2], especially for complex design configurations in the stepped wedge design such as unbalanced cluster size and non-uniform allocations of clusters to the

switch moments. Furthermore, within-cluster correlation structures and the value of the variance components are usually unknown. For that, reasonable guesses should be made from literature of similar trials. Otherwise, sensitivity analysis of different correlation structures and variance components values should be conducted.

Overall, the three elements should not be considered separately but rather as a whole. They are interconnected with each other. Decisions made on one element will influence the choice for the others. For instance, different design parameter choices requires different statistical methods such as a correlation structure, and will therefore change the required sample size substantially. Conversely, sometimes, it is necessary to modify the randomization unit or the design parameters to satisfy the restrictions of the sample size. Furthermore, there are much more aspects of the trial that need to be taken into careful consideration beside the three elements and factors mentioned above. For example, in Chapter 2, we also examined the double blind and informed consent problems. Since stepped wedge design is relatively new to the community, it is crucial to critically evaluate the decisions made on each aspect and have an objective and logical reasoning for that. A common pitfall is to treat the stepped wedge design based on the intuitions developed from classical parallel group design. Moreover, some variants of the stepped wedge design have been used in some of the trials in literature. [6, 8, 10, 11, 14]

7.2.2. STATISTICAL ANALYSIS

The key point presented in Chapter 3 about the statistical method for the stepped wedge design is to incorporate the design structure into the analysis. This structure includes the time-dependent nature of the treatment and the interplay between treatment effect and period. For example, to study patients' psychological difference between the care as usual follow-up protocol and the CEA-based follow-up protocol in Chap-

ter 6, an ANOVA-type of contrast was used. That is, a (generalized) linear mixed model was fitted to the data and the marginal mean responses at different blocks were estimated. Afterwards, assuming that the variance between the blocks could be explained by differences in patients, follow-up protocols and periods, treatment effect and period effect were obtained by linear combinations of these block means. The limitation is that treatment and period interactions could not be assessed. Although, it was explicitly assumed that observations from patients with two rounds both being under the intervention were different than the observations obtained from patients who were under the control at the first period and had switched to the intervention at the second time point. On the other hand, when studying the detection probability of the recurrences in Chapter 4, a marginal model with generalized estimating equation was used, exemplifying an alternative approach. Indeed, not only the design structure but also the correlation structures of the data need to be considered. For cross-sectional stepped wedge design, this usually means correlations within clusters and for longitudinal stepped wedge design it also means correlations within the same patient (sometimes correlation between patients from the same clusters as well). Chapter 4 and 5 showed one of the approaches to handle clustered survival times. In both chapters, a conditional model was used conditioning on the hospitals. However, there are several alternatives. A frailty model can be used for the conditional model approach. On the other hand, marginal approach with sandwich-type estimators can also be used. The two approaches are similar to the situation with the mixed model approach in Chapter 6 and the marginal model approach in Chapter 4, respectively.

7.3. GENERALIZATION AND FUTURE STUDIES

7.3.1. THE UNIDIRECTIONAL SWITCH DESIGN

In this thesis, the investigation is solely focused on the stepped wedge design. Essentially, the stepped wedge design belongs to a broader type of design called the unidirectional switch design. [17] Different switching moments in the stepped wedge design will lead to different patterns of one directional switching, thus explaining the name unidirectional switch design. If we further add switching at the beginning of the trial (a pure intervention switching pattern), and switching at the end of the trial (a pure control pattern), the collection of all the switching patterns forms the elements of the unidirectional switch design. All designs that only use these patterns are considered special cases of the unidirectional switch design. That includes the stepped wedge design as well. Another example of the unidirectional switch design is the delayed start design [7] which is frequently used in the development of drugs for Alzheimer's or Parkinson's disease for the purpose of demonstrating the disease modifying effect of a new drug. The delayed start design starts with a traditional parallel group design, but at a certain time point (part of) the control group switches to the new treatment. In addition, parallel group design can also be considered as a special case of the unidirectional switch design only with the pure control and intervention patterns.

It would be fruitful to further explore the unidirectional switch design. Because any general properties, such as sample size calculation, of the unidirectional switch design can be applied to multiple restricted forms such as the stepped wedge design. Such general framework, also provides flexibilities to study the differences and similarities between different designs which would lead to better understanding of when and why a particular design should be chosen. In addition, the unidirectional switch design solves the issues on treatment-period interactions with

the inclusion of the pure treatment and control patterns. The unidirectional switch design is a relatively new concept that generalizes some of the existing designs. Little work has been done for the unidirectional switch design. But the existing ones have already shown to be of practical value. [9, 17] For instance, it has been demonstrated that the general form of the unidirectional switch design (a design that includes all the switching patterns) could be more powerful than the stepped wedge design in terms of estimation efficiencies, and therefore requires smaller sample size. [9]

7.3.2. RANDOM SWITCH AND DYNAMIC TREATMENT

In the stepped wedge design, switching moments are randomized a priori and are independent of any characteristics of the patients. From a random switch perspective, stepped wedge design can be considered as randomly assign treatment to the participants among whom that have not been exposed to the intervention at each step. For example, consider a stepped wedge design with three clusters and three switch moments. From the traditional perspective, the three switch moments are randomly allocated to the three clusters. This is equivalent to randomly select one of the clusters with equal probability to receive the intervention at the first period, then randomly select one of the remaining two clusters with equal probability at the second period to receive the intervention, and at last assign the intervention to the last remaining cluster at the third period. The probability of each clusters being assigned to each pattern is still the same as a random allocation of the switch patterns to the three clusters upfront. This relates to the situation when switching to the new intervention is still at random but the probability depends on other factors such as patients' illness and the judgment of the doctors which is commonly seen in real life. For example, patients with much more severe symptoms have higher probability of receiving prescriptions from the doctor compared to

the healthier ones. Such phenomenon is quite prevalent in observational studies as well. The main problem is that the exchangeabilities of the patients can no longer be assured by the randomization procedure as in the stepped wedge design. Therefore, stepped wedge design can be considered as the randomized controlled trial counterpart of the random switch problem in the observational studies.

Another extension can be made from two treatments comparison to the comparison of treatment sequences. The latter is called dynamic treatment regime in the literature, which is a set of decision rules at different periods or stages of the disease indicating what treatment should be provided to the patients. [15] Comparisons of different dynamic treatment regimens plays an important role in personalized medicine since it generalizes personalization to time-varying treatment settings. [5] At different phases of the disease, choices of different treatment options are tailored to each patient based on the characteristics of that particular patient and the evolving information from the past. And studies of dynamic treatment regime can be informative to the evidence-based decision making procedure.

Overall, it is not a trivial task to analyze data from a stepped wedge design. Prudent considerations of the design during the planning phase of the trial and careful investigation on the statistical analysis methods are two important necessities to ensure the quality of the data and the fidelity of conclusions drawn from the trial. Nevertheless, the stepped wedge design is a very promising trial design option that has a lot of potentials to be further extended and generalized to more complicated situations.

REFERENCES

[1] Aalen O, Borgan O, Gjessing H (2008) Survival and event history analysis: a process point of view. Springer Science & Business Media

[2] Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ (2015)

Sample size calculation for a stepped wedge trial. Trials 16(1):354

[3] Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, et al (2015) Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials 16(1):1

[4] Brown CA, Lilford RJ (2006) The stepped wedge trial design: a systematic review. BMC Med Res Methodol 6(1):1

[5] Chakraborty B, Murphy SA (2014) Dynamic treatment regimes. Annual review of statistics and its application 1:447–464

[6] Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR (2015) Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials 16(1):352

[7] D'Agostino Sr RB (2009) The delayed-start study design. N Engl J Med 361(13):1304–1306

[8] Fatemi Y, Jacobson RM (2015) The stepped wedge cluster randomized trial and its potential for child health services research: a narrative review. Academic pediatrics 15(2):128–133

[9] Girling AJ, Hemming K (2016) Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Stat Med 35(13):2149–2166, DOI 10.1002/sim.6850

[10] Handley MA, Schillinger D, Shiboski S (2011) Quasi-experimental designs in practice-based research settings: design and implementation considerations. The Journal of the American Board of Family Medicine 24(5):589–596

[11] van den Heuvel ER, Zwanenburg RJ, van Ravenswaaij-Arts CM (2014) A stepped wedge design for testing an effect of intranasal insulin on cognitive development of children with Phelan-McDermid syndrome: A comparison of different designs. Stat Methods Med Res p 0962280214558864

[12] de Hoop E, van der Tweel I, van der Graaf R, Moons KG, van Delden JJ, Reitsma JB, Koffijberg H (2015) The need to balance merits and limitations from different disciplines when considering the stepped wedge cluster randomized trial design. BMC Med Res Methodol 15(1):1

[13] Hussey MA, Hughes JP (2007) Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 28(2):182–191

[14] Mdege ND, Man MS, Taylor CA, Torgerson DJ (2012) There are some circumstances where the stepped-wedge cluster randomized trial is preferable to the alternative: no randomized trial at all. response to the commentary by Kotz and colleagues. J Clin Epidemiol 65(12):1253

[15] Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. Mathematical Modelling 7(9-12):1393–1512

[16] Verberne CJ, Nijboer CH, de Bock GH, Grossmann I, Wiggers T, Havenga

K (2012) Evaluation of the use of decision-support software in carcino-embryonic antigen (CEA)-based follow-up of patients with colorectal cancer. BMC Med Inform Decis Mak 12(1):14, DOI 10.1186/1472-6947-12-14

[17] Zhan Z, de Bock GH, van den Heuvel ER (2017) Statistical methods for unidirectional switch designs: Past, present, and future. Stat Methods Med Res p 0962280216689280

SAMENVATTING

Binnen de geneeskunde wordt het gerandomiseerde gecontroleerde klinische onderzoek beschouwd als de gouden standaard voor het bepalen van de effectiviteit van een nieuwe interventie. In een gerandomiseerd gecontroleerd klinisch onderzoek worden patiënten random toegewezen aan verschillende behandelingen en worden de uitkomsten tussen de behandelgroepen vergeleken. De aanname is dat groepen die door middel van random toewijzing tot stand zijn gekomen, vergelijkbaar zijn wat betreft samenstelling van patiënten, waardoor verschillen in uitkomsten tussen groepen verklaard kunnen worden door de verschillen in behandeling. Binnen het gerandomiseerde gecontroleerde klinische onderzoek zijn er verschillende soorten onderzoeksdesigns, waaronder het bekende parallel group design, het cross-over design en het factorial design. Eén van de mogelijkheden binnen gerandomiseerd gecontroleerd onderzoek is het stapsgewijs (sequentieel) uitrollen van een nieuwe behandeling om deze vervolgens te vergelijken met een controle behandeling. In tegenstelling tot het klassieke parallel group design, waarbij verschillende behandelingen op hetzelfde moment worden toegewezen aan verschillende (clusters) patiënten, ontvangen (clusters) patiënten bij een stapsgewijze uitrol altijd beide behandelingen, waarbij de controle altijd voorafgaat aan de interventie. Dit biedt mogelijkheden om naast tussen- patiënten of tussen-cluster vergelijkingen ook binnen- patiënten of binnen-cluster vergelijkingen te doen. Bij de binnen- patiënten of binnen-cluster vergelijkingen kunnen patiënten en clusters als hun eigen controle worden beschouwd. Echter, het effect van de behandeling kan worden vertekend door het onderliggende tijdseffect, Een voorbeeld van zo'n design

is het zogenaamde stepped wedge design. Een stepped wedge cluster gerandomiseerd design is een speciale vorm van gerandomiseerd klinisch onderzoek en kan worden beschouwd als een mix van een parallel group design en een cross-over design. De basis van dit klinische studiedesign is dat de interventie stapsgewijs in de diverse clusters wordt gestart, en dat het moment van starten door het lot (dus random) wordt bepaald. Dit ontwerp lijkt vooral geschikt voor interventies waarvan de efficacy min of meer vast staat terwijl de effectiviteit nog een punt van onderzoek is.

Het stepped wedge design is populair onder epidemiologen en medisch onderzoekers vanwege de flexibiliteit en efficiëntie. Echter, zoals wel vaker zijn de voor- en nadelen van dit design in de praktijk meestal niet zo duidelijk als in theorie. Gedetailleerde en zorgvuldige overwegingen zijn dan ook nodig om de voor- en nadelen van dit design goed te beoordelen in vergelijking met andere designopties. Het plannen en uitvoeren van een stepped wedge design is uitdagend. In vergelijking met het klassieke parallel group design, moeten er meer keuzes worden bij het opzetten van een stepped wedge design studie. Een paar voorbeelden van keuzes zijn: het aantal patiënten, het aantal clusters, het aantal meetmomenten, het aantal switchmomenten en het aantal clusters per switchmoment. Al deze keuzes spelen een belangrijke rol in de opzet van een studie met het stepped wedge design. Daarnaast kunnen niet alle standaard methoden voor data-analyse direct toegepast worden op gegevens die zijn verzameld in het stepped wedge design, en voor bepaalde problemen in de analyse is nog geen oplossing bekend. Een voorbeeld van een stepped wedge design is de CEA-Watch studie. De CEA-Watch studie is een prospectieve gerandomiseerde gecontroleerde studie die in de periode 2010 tot 2012 in 11 Nederlandse ziekenhuizen werd uitgevoerd. De CEA-Watch studie kan worden gezien als een pragmatische trial. In dit onderzoek werd een intensief follow-up protocol vergeleken met een de gebruikelijke follow-up zoals beschreven in de landelijke richtlijn. In

de intensieve follow-up schema werd in de eerste drie jaar na diagnose elke twee maanden het CEA gemeten. In jaar 4 en 5 werd het CEA elke drie maanden gemeten. Jaarlijks werd een CT scan gedaan. Alleen bij een afwijkende bevinding van het CEA werd meer beeldvorming aangevraagd. In de gebruikelijke follow-up werd het CEA minder frequent gemeten (één tot twee keer per jaar) en werd er meer en vaker beeldvorming gedaan. Er werden 3223 patiënten in dit onderzoek ingesloten en 243 recidieven gediagnosticeerd. De belangrijkste conclusies zijn dat er met de geïntensiveerde follow-up strategie een hoger percentage behandelbare recidieven werden gevonden, en dat de recidieven die door screening warden gevonden een betere overleving hadden dan de recidieven die op basis van klachten waren vastgesteld. Patiënten hadden hoge verwachtingen van de geïntensiveerde follow-up maar waren ook enigszins bezorgd over het vooruitzicht van het vaak krijgen van een bloedtest. Op het tweede meetmoment waardeerden ze de geïntensiveerde follow-up meer dan op het eerste meetmoment. Wij zagen geen statistische significante nadelige effecten van de geïntensiveerde follow-up op de houding van patiënten ten aanzien van follow-up en psychologische functioneren. Als patiënten langer in de geïntensiveerde follow-up zaten werd hun waardering voor deze vorm van follow-up groter.Het doel van dit proefschrift, was het onderzoeken van praktische problemen die men tegen kan komen bij het uitvoeren van een stepped wedge design studie aan de hand de CEA-Watch trial, zoals de praktische implementatie van het ontwerp en de toepassingen van statistische methoden voor een stepped wedge design.

In hoofdstuk 2 zijn de sterke en zwakke kanten van het stepped wedge design besproken aan de hand van ons specifieke voorbeeld, de CEA-Watch trial. Dit hoofdstuk laat zien dat niet alle voordelen van het stepped wedge design aanwezig waren. Gelukkig bleef een groot deel van de voordelen van het stepped wedge design overeind in de CEA-Watch trial, terwijl de nadelen van het design minder problematisch bleken. Een be-

langrijk voordeel van het stepped wedge design in de CEA-Watch trial was de motivatie van patiënten en dokters om deel te nemen aan de studie omdat uiteindelijk iedereen de interventie (de geïntensiveerde follow-up) zou ontvangen. Dit had tot gevolg dat we in deze studie een relatief groot aantal patiënten konden includeren. Dit weegt, samen met de flexibiliteit van het stepped wedge design, op tegen de nadelen van het stepped wedge design zoals bijvoorbeeld de complexe statistische analyses. In de CEA-Watch trial, waarin het effect van intensivering van routine follow-up werd bestudeerd, is het van belang dat de onderzoeksperiode voldoende lang is. Dit geldt voor elkt type onderzoeksdesign. Door toepassing van het stepped wedge design krijgt het onderzoek, in vergelijking met andere soorten designs dan ook niet voor een langere onderzoeksperiode, ook niet als er sprake is van een longitudinale studieopzet. Natuurlijk moeten bij het ontwerpen van een trial altijd rekening gehouden worden met de specifieke trial en houden algemene principes, zoals voldoende lange onderzoeksperiode en vertekening door gebrek aan dubbelblindering, niet altijd stand in de praktijk. Maar in het algemeen kan gesteld worden dat het stepped wedge design een redelijk alternatief is voor het parallel groep design, zeker wanneer er sprake is van een grootschalige pragmatische onderzoek zoals bijvoorbeeld de CEA-Watch trial.

De statistische analyse van data verkregen met een stepped wedge design is ingewikkeld om verschillende redenen. Allereerst is de analyse complex vanwege de stapsgewijze introductie van de interventie. In hoofdstuk 3 werd door simulaties aangetoond dat het negeren van tijdseffecten, wanneer deze wel aanwezig zijn, kan leiden tot onzuivere schattingen van het behandelingseffect, en dat het geschatte behandelingseffect niet meer kan worden geïnterpreteerd als een gemiddeld behandelingseffect. Verder is het moeilijk om de interactie tussen tijd en behandeling te schatten omdat patiënten alleen in het begin van de studie aan de controle behandeling worden blootgesteld. In de literatuur is de meest toegepaste statistische

methode voor de analyse van data in een stepped wedge trial, het linear mixed model, zoals ontwikkeld door Hussey en Hughes. Echter, in deze analyse veronderstelt men een constant behandelingseffect gedurende de perioden. Dit blijkt niet altijd terecht te zijn. Tevens is het niet eenvoudig om de interactie van periode en behandeling in deze benadering op te nemen. Dit probleem wordt in hoofdstuk 3 geïllustreerd. Daarnaast wordt het clusterniveauanalyse van het stepped wedge design zelden besproken in de literatuur. Daarom stellen we in hoofdstuk 3 voor om meta-analyse technieken te gebruiken voor het analyseren van stepped wedge designs op cluster niveau wanneer tijdseffecten niet aanwezig zijn. Meta-analyse is een sterke kandidaat, aangezien het de mogelijkheden biedt om de heterogeniteit van de behandeling te beoordelen. Daarnaast zijn er in meta-analyse literatuur goed ontwikkelde theorieën om verschillende soorten uitkomsten te behandelen.

In de CEA-Watch trial, beschouwden we de ziekte van patiënten als een illness-death proces. Healthy state van het illness-death proces verwijst naar de ziektevrije toestand van patiënten. Zodra er recidieven werden gedetecteerd, werd de patiënt geplaatst in de ziekte toestand van het proces. De sterfte status van het proces heeft betrekking op het overlijden van de patiënten met of zonder recidief metastase. In hoofdstuk 4, de primaire uitkomsten waren de waarschijnlijkheid op detectie van recidieven en de tijd-tot-detectie door de follow-up protocol. Dit hoofdstuk was vooral gericht op het effect van follow-up protocol op de kans van overgang van de ziektevrije toestand naar de ziekte toestand. In hoofdstuk 5, de primaire uitkomsten waren de lange termijn overall survival tijd en ziektevrij survival tijd van de patiënten. In dit hoofdstuk werd het effect onderzocht van de geïntensiveerde follow-up protocol over de overgang van de ziekte toestand tot de overlijden toestand. Deze twee hoofdstukken waren twee studies over de twee overgangen in het één proces. Als alternatief kan de analyse worden uitgevoerd met behulp van een geavanceerder

multi-state model dat zowel het illness-death proces als de tijdsafhankelijke behandelingsschakelregeling omvat. Zonder in al te veel technische details te treden, zijn de gepresenteerde analyses in hoofdstuk 4 en hoofdstuk 5 voldoende voor het beantwoorden van de klinisch relevante vragen. In dit onderzoek hebben we niet alleen klinische uitkomsten bekeken, we hebben ook door patiënten gerapporteerde uitkomsten bekeken. Hoofdstuk 6 illustreerde de analyse voor gegevens uit vragenlijsten. Een moeilijkheid hierbij was dat de uitkomsten op slechts twee verschillende meetmomenten werden gemeten. Op het eerste meetmoment waren sommige ziekenhuizen al overgegaan op de nieuwe interventie, terwijl andere ziekenhuizen nog altijd standaardzorg volgens de Nederlandse richtlijn aanboden. Op het tweede meetmoment waren alle ziekenhuizen overgegaan op de interventie. De aanpak die we gebruikten in hoofdstuk 6 was om het linear mixed model te parametriseren, rekening houdend met de structuur van het studie. Hierna schatten we het behandeleffect gebruikmakend van combinaties van schattingen van het model. Het was van essentieel belang om rekening te houden met de tijdseffecten in de analyse, aangezien het zeer waarschijnlijk is dat de mening van patiënten over het follow-up protocol zullen veranderen over tijd. Retrospectief gezien zou het de voorkeur hebben om meer metingen in verschillende fasen van het onderzoek te hebben. Echter, gezien de tijd tussen elke meting en de tijd die nodig is om de vragenlijsten per post te versturen en te verzamelen, werd het niet mogelijk geacht meer meetmomenten toe te voegen.

Het geheel in aanmerking genomen maakt duidelijk dat analyse van data verkregen met een stepped wedge designed studie geen simpele opdracht is. Het kritisch overwegen van het onderzoeksdesign in de planningsfase en het grondig bestuderen van de statistische analysemethoden zijn twee onmiskenbaar belangrijke factoren voor de kwaliteit en de betrouwbaarheid van het onderzoek. Toch is het stepped wedge design

een veelbelovend design dat goede mogelijkheden biedt voor verdere uitbreiding om ook pure controle- en behandelarmen op te nemen en generalisatie naar een complexere situatie zoals dynamische behandeling en gerandomiseerde schakelen.

ACKNOWLEDGEMENTS

It took me four years of time to complete this thesis. During this exciting journey, I have been blessed with the accompany of many great persons. Without their inspiration, guidance, help, and support, I would not have been able to come this far. I take this opportunity to express my gratitude towards them.

First and foremost, my deepest gratitude goes to my supervisor Prof. Geertruida H. de Bock. Not only you have taught me the attitude of criticality, positive-ness, and responsibilities during my growth as an academic researcher, but also have helped me to become a better person with your tireless care and enormous understanding. It is great debt that I owe you and no single word suffice to express the appreciation.

I extend my appreciation to my second promotor, Prof. Edwin R. van den Heuvel who has shown and pointed out a direction of a life-long path for me and anchored a theme on the prologue of a new start. Thank you for your support through out the bumpy ride and I look forward to the day to repay your favour with the most beautiful language of mathematics.

Much of the work of this thesis is the results of collaboration with the CEA-Watch project members, to whom I wish to express my appreciation as well. First of all, Prof. Theo Wiggers, it was an amazing experience to just listen to you talking about colorectal cancer. Your thoroughness, and willingness to share your expertise are much appreciated. And I would also like to thank your insightful advice and enjoyable discussions. Dr. Charlotte J. Verberne, it has been delightful to work with you and I am very proud on our joint work which covers substantial part of this thesis. I also like to pass my gratitude to Dr. Irene Grossmann, Prof. Adelita V.

Ranchor, Fabian J. van der Sluis and all other members of the project for their contributions to the project. To Femke Oppers, Anne Marye de Jong who allowed me to work on their master thesis.

I am also thankful to prof. H. M. Boezen, prof. G. J. P. van Breukelen, and prof. G. Beets for being part of my reading committee, for their time to read this thesis and for providing me with valuable feedbacks.

To prof. N. Balakrishnan, prof. Wenli Lu, Dr. Sandra Geurts and Anne Aarts for your hospitalities during my visit to McMaster University, Tianjin Medical University, and Radboud University Nijmegen, respectively; To all the new friends and colleagues in Eindhoven as well, as they have welcomed me as one of their own members during the last year of my Ph.D study; To Ossama Almalik and Dr. Nazanin Nooraee, I cherish every minute we spent together discussing all sort of statistical and non-statistical problems.

To colleagues and friends at University Medical Center Groningen. To Dr. Marcel W. J. Greuter, for introducing me to the world of academia; Hans Burgerhof and Dr. Sasha la Bastide-van Gemert, two amazing statisticians I have ever met, for unselfishly sharing your knowledge about statistics with me; Dr. Tallitha F. Feenstra, for the detailed and insightful comments and discussions during the collaboration of the breast cancer modeling project; Anne Looijmans and Qi Cao, for being such awesome office mates; Oncological epidemiology unit members, Dr. Janet de Vos, Dr. Rositsa G. Koleva-Kolarova, Xuan Anh Phí, and all other members for your kindness, help, and the fruitful discussions from which I have learned a lot about breast cancer screening; Dr. Daan "Coach Daan" Reid and his family, Dr. Douwe Postmus, Dr. Gert van Valkenhoef, for the exciting "gaming night and weekends"; Leane Kuipers, Diana van der Plaats, Brechtsje Kingma, Kim de Jong, and all lunch group members for the enjoyments during the lunch time; Pepijn Vemer, your enthusiasm for statistics, Majong, and ties has never seized to amaze me; Roelian J. Geuze, Aukje van der Zee, and

Marijke Hanania, for their patient secretarial support and finance-related administrative support; and all colleagues I have come across with, whom I have valued no less.

To all my Chinese friends, especially Sun Zhe & Qin Jing, Suhai Zhang, Zetao Chen, Jieqiang Wei, Ze Li, Xin Liu, Kexin Wang, Qi Cao, Congchao Lv, Xiang Zeng, Yi Jiang & Ting Zhao, Bin Han & Haoxiao Zuo, Jun Li, Ming Li, Zhuoran Yin, Qi Wang, Liwen Zhang, and Huishuai Li, for making this memory even more unforgettable.

To my three lovely paranymphs, Anne Looijmans, Xuan Anh Phí and Bronislaw Abramiuc for all the jokes and laughter, all your help and all your accompany. Thank you and AIR HIGH FIVE! Special thanks to Anne Looijmans for spending enormous time on translating the Dutch summary and Anh Phí for helping with the defense when Anne is not available. For “my dear friend” Dada Broni and his ingenious opinions about life and people.

Last but not the least, I dedicate this thesis to my family, to whom I owed too much and would not have finished this thesis without their support and understanding.

CURRICULUM VITÆ

Zhuozhao ZHAN

10-11-1988 Born in Wenzhou, China.

EDUCATION

2007–2011 Bachelor of science
Central South Univerisity, Changsha, China

2011–2013 Master of science
University of Groningen, Groningen, The Netherlands

2013–2017 Ph.D candidate
University of Groningen,
University Medical Center Groningen,
Groninenge, The Netherlands
Thesis: Evaluation and analysis of stepped wedge design: Application to colorectal cancer follow-up
Promotor: Prof. dr. G. H. de Bock
Prof. dr. E. R. van den Heuvel

PUBLICATIONS

2014 [1] Z. Zhan, E. R. van den Heuvel, P. M. Doornbos, H. Burger, C. J. Verberne, T. Wiggers, and G. H. de Bock, *Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study*, Journal of clinical epidemiology 67, 454, 2014

2015 [2] R. G. Koleva-Kolarova, Z. Zhan, M. J. W. Greuter, T. L. Feenstra, G. H. de Bock, *Simulation models in population breast cancer screening: a systematic review*, The Breast 24, 354, 2015

[3] C. J. Verberne, Z. Zhan, E. R. van den Heuvel, I. Grossmann, P. M. Doornbos, K. Havenga, E. Manusama, J. Klaase, H. C. J. van der Mijle, B. Lamme, K. Bosscha, P. Baas, B. van Ooijen, G. Nieuwenhuijzen, A. Marinelli, E. van der Zaag, D. Wasowicz, G. H. de Bock, T. Wiggers, *Intensified follow-up in colorectal cancer patients using frequent Carcino-Embryonic Antigen (CEA) measurements and CEA-triggered imaging: Results of the randomized "CEAwatch" trial*, European Journal of Surgical Oncology, 41, 1188, 2015

[4] R. G. Koleva-Kolarova, Z. Zhan, M. J. W. Greuter, T. L. Feenstra, G. H. de Bock, *To screen or not to screen for breast cancer? How do modelling studies answer the question?*, Current Oncology 22, e380, 2015

2016 [5] Z. Zhan, G. H. de Bock, T. Wiggers, E. R. van den Heuvel, *The analysis of terminal endpoint events in stepped wedge designs*, Statistics in Medicine, 35, 4413, 2016

2017 [6] Z. Zhan, G. H. de Bock, E. R. van den Heuvel, *Statistical methods for unidirectional switch designs: Past, present, and future*, Statistical Methods in Medical Research, Online First: 0962280216689280, 2017

[7] F. J. van der Sluis, Z. Zhan, C. J. Verberne, A. C. M. Kobold, T. Wiggers, G. H. de Bock, *Predictive performance of TPA testing for recurrent disease during follow-up after curative intent surgery for colorectal carcinoma*, Clinical Chemistry and Laboratory Medicine, 55, 269, 2017

[8] C. J. Verberne, Z. Zhan, E. R. van den Heuvel, F. Oppers, A. M. de Jong, I. Grossmann, J. M. Klaase, G. H. de Bock, T. Wiggers, *Survival analysis of the CEAwatch multicentre clustered randomized trial*, British Journal of Surgery, 104, 1069, 2017

[9] Z. Zhan, C. J. Verberne, E. R. van den Heuvel, I. Grossmann, A. V. Ranchor, T. Wiggers, G. H. de Bock, *Psychological effects of the intensified follow-up of the CEAwatch trial after treatment for colorectal cancer*, PloS one, 12, e0184740, 2017